JAN 28 1980

EP4.2: P41

United States
Environmental Protection
Agency

Office of Air Quality
Planning and Standards
Research Triangle Park NC 27711

EPA-450/4-79-033
October 1979

Air

EPA

Procedures for Evaluating the Performance of Air Quality Simulation Models

COMPLETED

ORI

158

EPA-450/4-79-033

Procedures for Evaluating the Performance of Air Quality Simulation Models

by

M.J. Hillyer, S.D. Reynolds, and P.M. Roth

Systems Applications, Incorporated
950 Northgate Drive
San Rafael, California 94903

Contract No. 68-02-2593

EPA Project Officer: Russell F. Lee

Prepared for

U.S. ENVIRONMENTAL PROTECTION AGENCY
Office of Air, Noise, and Radiation
Office of Air Quality Planning and Standards
Research Triangle Park, North Carolina 27711

October 1979

i

This report is issued by the Environmental Protection Agency to report technical data of interest to a limited number of readers. Copies are available free of charge to Federal employees, current contractors and grantees, and nonprofit organizations - in limited quantities - from the Library Services Office (MD-35), U.S. Environmental Protection Agency, Research Triangle Park, North Carolina 27711; or for a nominal fee, from the National Technical Information Service, 5285 Port Royal Road, Springfield, Virginia 22161.

This report was furnished to the Environmental Protection Agency by Systems Applications, Inc., 950 Northgate Drive, San Rafael, CA 94903, in fulfillment of Contract No. 68-02-2593. The contents of this report are reproduced herein as received from Systems Applications, Inc. The opinions, findings, and conclusions expressed are those of the author and not necessarily those of the Environmental Protection Agency. Mention of company or product names is not to be considered as an endorsement by the Environmental Protection Agency.

Publication No. EPA-450/4-79-033

ACKNOWLEDGMENT

The authors would like to express their appreciation to Stan Hayes for many helpful discussions and useful comments during the preparation of this report.

BLANK PAGE

CONTENTS

DISCLAIMER ii

ACKNOWLEDGEMENT. iii

LIST OF ILLUSTRATIONS. vii

LIST OF TABLES ix

LIST OF EXHIBITS xi

I INTRODUCTION. 1

- A. Purpose of This Study 2
- B. Structure of This Report. 3

II REVIEW OF PREVIOUS MODEL EVALUATION EFFORTS 5

III KEY ISSUES IN THE EVALUATION OF AIR QUALITY 9

IV A PROCEDURE FOR EVALUATING MODEL PERFORMANCE. 17

- A. Problem Specification and Model Selection 20
  - 1. Setting the Context of Model Usage. 20
  - 2. Selection of a Model. 22
- B. Planning the Model Evaluation Study 23
  - 1. Determination of the Need for Model Evaluation. . . . 23
  - 2. Development of a Conceptual Evaluation Plan 26
  - 3. Examination of Existing Data Bases. 37
  - 4. Assessment of Data Needs. 43
  - 5. Assessment of the Need To Collect Additional Data. 64
  - 6. Specification of Performance Standards and Measures. 65
- C. Identification of the Scope and Requirements of Model Evaluation. 65
- D. Performance of the Model Evaluation 67
  - 1. Adapting the Model for Use in the Study 67
  - 2. Gathering, Assembling, and Formatting the Required Data 71
  - 3. Exercising the Model. 72

4. Analyzing the Results of the Evaluation 72
5. Assessing the Need To Perform Further Evaluation. 75
6. Evaluating the Adequacy of the Model. 75

E. Evaluation for Screening Applications 76

F. Perspective 77

V RECOMMENDATIONS 79

A. Institutional Needs 79

B. Areas for Technical Development 81

C. Documents To Be Compiled. 82

D. Summary 83

APPENDIX: SUMMARY OF PREVIOUS EVALUATION STUDIES. 85

REFERENCES 141

ILLUSTRATIONS

IV-1 Simplified Flow Diagram of Tasks in the Model Evaluation Process 18

IV-2 Comparison of Calculated Ground Trajectory with Observed Tetroon Trajectory for the Los Angeles Basin. 49

IV-3 Surface CO Concentrations Calculated for the San Francisco Bay Area at 1400 on 26 July 1973 Using the LIRAQ Air Quality Model at Different Grid Sizes. 69

A-1 A Typical Cumulative Frequency Distribution of 24-Hour-Average SO_2 Concentrations Measured and Calculated Using the CRSTER Model. 104

A-2 J. M. Stuart Plant Cumulative Frequency Distribution for One-Hour-Average SO_2 Concentrations at All Stations. . . . 105

A-3 Measured SF_6 Concentrations and Predictions Using PTMTP for Stability Categories B Through F 108

A-4 Scatter Diagram Comparing Observed and Calculated Concentrations Obtained from the Standard Gaussian Plume Model Under Slightly Unstable Conditions (C Stability) for the High Source at the Western Kraft Corporation 109

A-5 Comparison of CO Measurements and Predictions from Four Models Examined by Maldonado and Bubbin. 114

A-6 Time Variations over All Stations of Observed One-Hour-Average Ozone Concentrations and the Corresponding Predictions Obtained from the SAI Airshed Model. . . . 120

A-7 Predictions of the SAI Airshed Model Compared with Estimates of Instrument Errors for Ozone (Date for 3 Days, 9 Stations, Daylight Hours). 121

A-8 Comparison of Measured and Predicted 25-Day-Averaged SO_2 Concentrations (1-26 February 1965). 128

A-9 Comparison of Predictions and Measurements of SO_2 Concentrations in St. Louis Reported by Shir and Shieh. 130

A-10 Comparison of Predicted and Measured Two-Hour-Averaged SO_2 Concentrations at Each Monitoring Station in St. Louis. 131

A-11 Comparison of Predicted and Measured Two-Hour-Averaged Frequency Distributions of SO_2 Concentrations According to Wind Sectors. 132

A-12 Comparison of Frequency Distributions of Predicted and Measured Two-Hour-Average SO_2 Concentrations According to Ambient Temperature Range 133

A-13 Comparison of Frequency Distributions of Predicted and Measured Two-Hour-Average SO_2 Concentrations for Three Stations 134

A-14 Statistical Quantities Calculated for Nine Air Quality Simulation Models. 138

TABLES

IV-1 Possible Data Requirements of Air Quality Models 30

IV-2 Possible Hardware Requirements for Data Collection . . . 36

IV-3 Number and Type of Daily Meteorological Measurements in 15 U.S. Cities in 1977. 41

IV-4 Methods Used for the Preparation of Emissions Inventories in 15 U.S. Cities in 1977. 42

IV-5 Number of Stations Performing Routine Air Quality Sampling in 15 Major Cities in the United States 45

IV-6 Estimated Meteorological Data Requirements for Evaluation of a Large Air Quality Simulation Model in a Hypothetical Urban Area 51

IV-7 Levels of Detail in Data Used as Input to Photochemical Air Quality Simulation Models. 58

IV-8 Values and Sources of Errors in Meteorological Measurements 61

IV-9 Uncertainties in Measurements of Pollutant Concentrations. 62

IV-10 Model Performance Measures and Standards 66

A-1 Models Considered in Model Evaluation Studies Described in This Appendix 87

A-2 Statistical Summary of Observed Two-Hour SO_2 Concentrations (in $\mu g/m^3$) for St. Louis Stations and Concentrations Calculated Using a Gaussian Plume Model. 91

A-3 Statistical Summary of Observed One-Hour SO_2 Concentrations (in $\mu g/m^3$) for Chicago Stations and Concentrations Calculated Using a Gaussian Plume Model. 92

A-4 Comparison of Error Distributions for Two-Hourly St. Louis and Hourly Chicago Validation Calculations . . 93

A-5 Observed, Predicted, and Observed Minus Predicted Concentrations by Wind Speed Class for St. Louis Data 94

A-6 Observed, Predicted, and Observed Minus Predicted Concentrations by Wind Speed Class for Chicago Data 94

A-7 Statistics for Long Term Average Predictions of SO_2 95

A-8 Summary of Accuracy of Sampling Intervals for Estimating Distribution of Predicted Concentrations over a Season. 96

A-9 Comparisons of Measured and Predicted One-Hour-Average SO_2 Concentrations in New York City. 99

A-10 Model Comparisons for Annual Mean SO_2 Concentrations Using New York City Data. 100

A-11 Model Comparisons for Annual Mean Particulate Concentrations Using New York Data 101

A-12 Correlation Between Observed SF_6 Concentrations and Concentrations Calculated Using PTMTP, by Stability Class. 108

A-13 Parameter Combinations Employed in the Prahm and Christensen Study and a Summary of the Results 111

A-14 Statistical Results from the Maldonado and Bullin Study 113

A-15 Occurrence of Corresponding Levels of Predicted and Observed Ozone 122

A-16 Statistical Measures Used in Verification of the LIRAQ Photochemical Model 124

A-17 Statistics for the LIRAQ Model Evaluation Study. 125

EXHIBITS

IV-1 Air Quality Monitoring Required by the EPA for Various Pollutants 44
Vario
IV-2 Performance Specifications for Automated Measurement Methods. 63

I INTRODUCTION

Over the past several years there has been a significant increase in the use of air quality models. This increase has stemmed primarily from governmental regulations, notably the Clean Air Act of 1967 and amendments to it in 1970 and 1977. Air quality models are potentially applicable to the evaluation of State Implementation Plans, Prevention of Significant Deterioration, New Source Reviews, and Indirect Source Reviews. In addition, they can be used in the preparation of Environmental Impact Statements and in support of various litigation issues. In these applications the pollutants of main interest are those for which air quality standards have been developed, including particulates, SO_2, CO, oxidants, and NO_2. Other pollutants, such as hydrocarbons and NO, must be considered in some applications because they are involved in the formation of pollutants regulated by standards. The issues and the pollutants mentioned above pose many different requirements for air quality modeling, and the standards themselves add further requirements. They necessitate pollutant concentration predictions over averaging periods ranging from one hour to one year. Many models have been developed to deal with the different issues, pollutants, and standards in different physical settings, but no single model is suitable for all applications.

Given this variety of models, there is a need for an adequate understanding of the performance of an air quality model, both relative to that of other models and in an absolute sense in a particular application. This understanding is required to enable the user to choose an appropriate model and to ensure that it performs adequately in the intended application. In the past, model performance was usually evaluated by the model developers in the course of an application or a special evaluation study. As model evaluation becomes an integral part of air quality studies and as the number of these studies increases, there will be an increased need for uniform procedures for conducting model evaluation studies. At the present time, uniform procedures do not exist.

A. PURPOSE OF THIS STUDY

Stimulated by the need for an objective framework for conducting evaluations of model performance, as expressed in the "Report to the U.S. EPA of the Specialists' Conference on the EPA Modeling Guideline" (Argonne, 1976), the Environmental Protection Agency (EPA) requested that Systems Applications, Incorporated carry out a study under Work Assignment 1 of Contract 68-02-2593. This work assignment calls for the development of evaluation procedures for both short-term and long-term models, for both area-wide and site-specific applications. In a companion study under the same work assignment, model performance measures and standards are being examined to aid in assessing the relative accuracy and usability of any model for various applications (Hayes, 1979).

The objective of this study is to lay the groundwork for a set of guidelines that can be implemented to aid in assessing performance characteristics of air quality models. In carrying out this work, we have:

> Reviewed previous model evaluation studies.

> Developed a general procedural framework for performing an evaluation study.

> Provided specific guidance, to the extent possible, with respect to the work required in each step of the performance evaluation procedure.

> Examined the role appropriate organizations should play in overseeing model evaluation studies.

> Identified gaps in present knowledge that limited our ability to provide more detailed guidance in this report, and presented recommendations for further work that will help to fill those gaps.

Because model evaluation has received relatively little systematic attention to date, we were able to identify several areas ripe for future investigation. The performance of these suggested studies will be essential to the success of the guidelines presented herein.

Some of the terminology used to describe the model evaluation process may require clarification. The phrases "model validation" and "model verification" are frequently used to designate the process of comparing model predictions with suitable observations. In this report, however, we avoid the terms "validation" and "verification" when referring to the evaluation of a model. Instead, we take a more general view and use the phrase "model performance evaluation," which we believe to be more indicative of the process that this report describes. The "validity" of a model is taken to be a concept defining how well model predictions would agree with the appropriate observations given a perfect specification of model inputs. That is, validity relates to the inherent quality of the model formulation. The term "verification" is reserved to describe a successful (or positive) outcome of the model evaluation process.

B. STRUCTURE OF THIS REPORT

Chapter II gives a brief overview of previous model performance evaluation studies; detailed descriptions of the results of these studies are given in the appendix. The objective of Chapter II is to provide some perspective on how model performance has been assessed in the past. In particular, we attempt to identify some of the weaknesses and inadequacies of previous efforts so that such shortcomings can be avoided in the proposed guidelines. The models employed in these investigations range from relatively simple Gaussian plume models to comprehensive numerical models suitable for studying photochemical air pollution in urban areas.

Based on our review of previous studies and further consideration of the model evaluation process, we identify in Chapter III some of the key questions and issues that need to be addressed in the model evaluation guidelines. This discussion is intended to set the stage for the more detailed exposition of the guidelines given in Chapter IV.

Chapter IV is devoted to the development of a step-by-step procedure for carrying out a complete model evaluation study. This procedure is segmented into four sequential phases: problem definition and model selection; preliminary planning efforts; formal definition of the evaluation plan; and implementation of the evaluation study. To the extent possible, we discuss the elements of each step in each phase, giving special consideration to the influence that a particular model or type of application might have on how each step is performed.

Chapter V concludes this report with a set of recommendations for future conduct of model evaluation studies. Specifically, we identify:

> A set of institutional recommendations toward a possible organization for ensuring proper evaluation procedures in modeling work.

> A series of recommended research studies necessary to assemble vital background information for model evaluation.

> A list of guidelines documents that should be available to those model users who have a need to evaluate a model.

It is our belief that implementation of these recommendations in conjunction with the guidelines presented in Chapter IV would yield a complete protocol for structuring a model performance evaluation study.

II REVIEW OF PREVIOUS MODEL EVALUATION EFFORTS

The need for evaluating the performance of air quality models is well recognized by model developers, and evaluation studies are often undertaken as a part of model development. In such an evaluation, the goal is to determine whether the given model adequately reproduces representative ambient air pollutant concentrations for the set(s) of conditions for which it was designed. "Representative" refers to concentration measurements that are commensurate with the model's predictions in terms of both temporal and spatial scales. The basic data requirements for a model evaluation study are twofold: the input data necessary to exercise the model and the monitoring data appropriate for checking the model's output. In this chapter, we review previous model evaluation efforts, concentrating on the methods used. The appendix describes the results of these studies in detail.

Two aspects of air pollutant concentration variations should be accounted for by an air quality model. First, it should correctly predict the amounts of pollutants that are formed (as indicated by pollutant concentrations). Second, it should account satisfactorily, as far as its formulation permits, for both spatial and temporal trends in the pollutant concentration fields. The accuracy with which these trends should be reflected depends somewhat on the contemplated application of the model. If it is to be used for assessing population exposure to peak pollutant impacts, accurate location and timing of the predicted peak would be desirable to take diurnal population shifts into account. In contrast, the model may be used to compute the peak impact from a point source or in an urban airshed to judge compliance with an air quality standard. In that case, some offset in space and time may be tolerable provided that the concentration itself is accurately predicted.

Discrepancies between observed pollutant concentrations and those predicted by a model can arise from several sources:

> Inaccuracy of the data input to the model.
> Inadequate or insufficient information to define or specify model inputs properly.
> Inadequacies and simplifying assumptions made in the model's formulation.
> Inaccuracy (due to both measurement error and nonrepresentativeness) of air quality data used to evaluate model predictions.
> Inaccuracies introduced by numerical solution procedures.

In a well planned evaluation study, all data inadequacies are minimized so that model inadequacies can be properly identified.

In the studies considered in this review, the models fell into two basic categories: Gaussian and numerical. Evaluation procedures were limited to comparisons between measured and calculated pollutant concentrations. In some studies, these comparisons were made qualitatively using visual techniques such as scatter plots; in others comparisons were quantitative using mathematical or statistical techniques to develop numerical figures of merit for model evaluation.

In our review, we did not attempt to include all of the evaluation studies carried out to date; rather, we examined some representative examples and the methods used by different workers to evaluate air quality models. The results of these previous efforts can be used to guide our development of a comprehensive and logically constructed model performance evaluation procedure.

Numerous air quality models are based on the Gaussian formulation. These models have been developed to cover a wide range of applications, from single sources to complete urban areas, and for averaging times from one hour to one year. Such models are widely used, particularly for assessing pollutant impacts from point sources.

Since all Gaussian models are based on the same dispersion theory, they should be equally accurate in terms of that aspect. However, other model components, such as the treatment of plume rise, introduce differences among models. A comparison of Gaussian model performance, classified by model type (e.g., single-source models, multisource models, climatological models), would provide valuable assistance in model selection. Unfortunately, no such comparison is available because no protocol exists for performing it. The evaluations of Gaussian models that have been carried out reflect the particular needs of the evaluators rather than the general needs of model users. The evaluations have varied widely in effectiveness, ranging from comparative studies of three or four models to brief studies of a single model under a restricted set of conditions. Both empirical and quantitative statistical methods have been used, but the lack of standardized methods precludes comparisons of study results. The model comparisons that can be derived from studies of more than one model are limited by the data base as well as by the performance measures used; generally that data base has been limited to available data. As we discuss later in this report, available data are usually unsatisfactory for model evaluation. Thus, data bases collected for the specific purpose of the given study are highly desirable.

Evaluation studies of numerical models generally suffer the same problems as the Gaussian model studies: noncomparable methodologies and inadequate data bases. Again, no absolute indicators of the performance of particular models exist because of the lack of appropriate performance standards against which to measure them. As in the Gaussian model case, most evaluation studies were undertaken using a preexisting data base that is inadequate for model evaluation.

Although many models have been subject to evaluative studies, no air quality model has been adequately verified. The main reason for this failure is the lack of standards against which to measure model performance. Moreover, no systematic procedures have been developed for model evaluation or interpertation of evaluation results. Consequently, studies cannot be compared with one another or assessed in absolute terms.

What little work has been carried out to investigate model evaluation methodologies has centered on assessing the statistical aspects of model performance measures.

Examination of the model evaluation studies considered in this chapter and in the appendix reveals that many of these efforts are undertaken by the model developer. Although model evaluation is an important and proper adjunct to model development, independent evaluation of air quality models may provide a more objective basis for discrimination. Also, it seems reasonable to require that the validity of a model be established according to some standard guidelines before the model is used. Proper verification of an air quality model will provide confidence that the model predictions are a reasonable representation of the possible effect of a particular future emissions pattern.

Later in this report, we discuss the contents of such a set of guidelines, dealing first with the general issues that must be addressed in a model evaluation effort and then with more specific recommendations regarding the components of such a study. An indispensable corollary to the formulation of guidelines for model evaluation studies is the establishment of model performance measures and standards. This problem is being addressed in a concurrent study and is the subject of a companion volume (Hayes, 1979).

III KEY ISSUES IN THE EVALUATION OF AIR QUALITY MODEL PERFORMANCE

Our review of previous studies, discussed in Chapter II, identified a number of issues that should be addressed in a model performance evaluation study. The purpose of this chapter is to delineate these issues and to provide the reader with an introductory perspective regarding their relevance. These issues form the basis for the step-by-step model evaluation procedure presented in the next chapter.

Model performance is influenced by two factors: First, the model treatments of various physical and chemical phenomena usually involve a number of approximations; second, the information provided as inputs is often subject to considerable uncertainty. To illustrate the first point, we consider photochemical air quality simulation models. Seinfeld (1977) shows that the governing equation for such models can be written:

$$\frac{\partial\langle\bar{c}_i\rangle}{\partial t} + \bar{u}\,\frac{\partial\langle\bar{c}_i\rangle}{\partial x} + \bar{v}\,\frac{\partial\langle\bar{c}_i\rangle}{\partial y} + \bar{w}\,\frac{\partial\langle\bar{c}_i\rangle}{\partial z} = \frac{\partial}{\partial x}\left(K_H\,\frac{\partial\langle\bar{c}_i\rangle}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_H\,\frac{\partial\langle\bar{c}_i\rangle}{\partial y}\right)$$

$$+ \frac{\partial}{\partial z}\left(K_V\,\frac{\partial\langle\bar{c}_i\rangle}{\partial z}\right) + R_i(\langle\bar{c}_1\rangle,\ldots,\langle\bar{c}_N\rangle,T) + \langle\bar{S}_i\rangle \quad , \qquad (1)$$

where

c_i = concentration of pollutant i,

K_H, K_V = horizontal and vertical diffusivities, respectively,

u, v, w = wind velocity components in the x, y, and z directions, respectively,

R_i = net rate of production of species i by chemical reactions,

S_i = net rate of emissions of species i.

The symbol "-" indicates the result of averaging the conservation of mass equation over an ensemble of flows, and the symbol "<>" indicates the result of averaging over a volume corresponding to the size of a grid cell (as needed to obtain a numerical solution).

To derive Eq. (1) from the more general mass conservation relationship, the following assumptions are necessary:

- > Molecular diffusion is negligible when compared with advective transport.
- > Turbulent transport can be represented by the eddy-diffusivity concept.
- > The influence of turbulent and subgrid-scale concentration fluctuations on chemical reaction rates can be neglected.

The neglect of molecular diffusion is generally acknowledged to be satisfactory in atmospheric applications (Seinfeld, 1977). Thus, the primary sources of inherent invalidity in Eq. (1) are the treatments (or neglect) of turbulence effects on transport and on reaction phenomena.

Model performance is also affected by the error associated with each of the model parameterizations and inputs. In photochemical models based on Eq. (1), the sources of these uncertainties include (Seinfeld, 1977):

- > Wind velocity components.
  - Uncertainties in available wind speed and direction measurements.
  - Inadequate number or nonrepresentative locations of measurement sites (especially aloft).
  - Approximations associated with wind field analysis techniques used to prepare model inputs.
- > Source emissions function.
  - Inaccurate specification of source location.
  - Inaccurate estimation of plume rise.

- Errors in emissions factors for stationary sources.
- Inadequate representation of actual driving characteristics in the federal mobile emissions test procedures.
- Errors in emission factors and vehicle miles traveled for mobile sources.
- Inaccurate characterization of temporal variations.

> Chemical reaction mechanism.
  - Omission or inadequate characterization of chemical reaction steps.
  - Uncertainties in measurement or specification of reaction rate constants.
  - Inaccurate characterization of temperature effects.
  - Uncertainties associated with categorizing species into reactive groups (such as paraffins, olefins, aromatics, etc.).

> Initial and boundary conditions.
  - Inadequate spatial characterization of the concentration field on the upwind boundary of the region.
  - Inadequate characterization of concentrations aloft.

> Numerical solution methodology.
  - Computational errors associated with the use of finite difference methods.
  - Computational errors associated with other numerical techniques.

In addition, we note that the air quality data employed to judge model performance are subject to error. Instrumental errors can occur, or the spatial character of the measurement may not be directly comparable to that of the model predictions (e.g., a point measurement may be compared to a spatially averaged model prediction). Thus, there are a number of sources of error in even the most sophisticated air quality models. Under nonideal conditions, the errors in predictions produced by relatively simple models (e.g., Gaussian) can be greater than those generated by more sophisticated techniques that attempt to provide a more adequate representation of the nonideal conditions or processes. The purpose of the model evaluation study is to test the entire model, including formulation, available aerometric and emissions data, and model input preparation procedures, to obtain some quantitative measure of model performance. Hilst (1978) suggests in a recent report the additional need to evaluate the performance of specific components of a model. For example, special studies could be performed to assess the plume rise algorithm included in a Gaussian point source model. These types of investigations will be especially important in the initial applications of a model because they may aid in the identification and rectification of inadequate treatments of atmospheric processes.

A model evaluation study consists of several tasks, many of which are interrelated. As the project team undertakes these tasks, a number of questions will arise. Some of these issues can be resolved by establishing performance evaluation guidelines, but others will have to be addressed by the user in light of the particular situation. Among the most important questions that may arise in an evaluative study are the following:

> How does the user select a suitable model?
> Under what circumstances must the user undertake a model evaluation study?
> What are the important elements of an evaluative study and how should it be planned?
> What are the resources available to the study team?
> What sources of model evaluation expertise are available to the user?
> What are the aerometric and emissions data requirements of the model?
> What are the appropriate measures and standards of model performance?
> What are the appropriate analyses of model results?
> When must the user mount a supplemental field measurement program?
> What should the user do if the performance of the model is found to be unsatisfactory?

In the remainder of this chapter we identify additional issues associated with these questions.

The identification of a suitable model for calculating air quality levels will be one of the first tasks to be performed by a model evaluation study team. To aid the inexperienced user, appropriate model selection guidelines that address the various types of models and their range of applicability (e.g., pollutants, averaging time and chemical effects, etc.) should be published. The user will need to give careful consideration to model selection to avoid, as far as possible, having to make extensive modifications to the model or even having to select another model.

To facilitate cost-effective and timely use of models, consideration should be given to identifying situations requiring model performance evaluation. For example, must the user always provide direct evidence of model performance? What is the value of previous evaluative experience, and when can it be cited by the user in lieu of conducting a new study? What assistance can the EPA or other model evaluation experts provide to the user in making these important decisions?

Once a decision to proceed with the evaluative effort has been reached, the project team must develop a comprehensive plan for the study. The guidelines should identify the various elements of such a study and detail what is required in each step.

Three types of resources will be required by the study team: financial, hardware (both aerometric measuring instruments and computers), and manpower. Key issues include the amount of funding required for the evaluative study and possible sources of financial assistance, the number of air quality and meteorological instruments and other associated hardware (in addition to that available from the routine monitoring network) required in a supplemental field monitoring program, the computer hardware needed to carry out the model calculations, and the types of technical expertise that the study team should include. It may be important to identify compromises that can be made to reconcile the desired resources with those actually available.

To estimate the required resources, the investigating team must have a clear understanding of the model's data requirements and their relation to the various influences on model performance, such as region size, terrain characteristics, emissions patterns, meteorological conditions, etc. If the available aerometric and emissions data base is not adequate to characterize some of the model inputs, the user will need some guidance in identifying alternative means for specifying these parameters.

Perhaps the best means for reconciling the data needs of a model with the available data base is to undertake a special field measurement program to provide aerometric and emissions data specifically for the model evaluation and subsequent application. Because the costs of such a program can be considerable, the study team should have sufficient information to enable them to effectively plan and implement such an effort. To assess model performance and results, measures and standards should be established that take into account the types of available models and their application. These measures and standards should be readily computed from model results. Their publication will help users to identify the pertinent measures and standards and to tailor the evaluation program to ensure that the standards are met.

Given a set of model predictions, the user must perform various analyses to assess model performance. An important element of the model evaluation guideline should be the identification of analyses that will help the user to characterize model performance as well as to uncover any significant biases or flaws in either the model formulation or its inputs.

The achievement of satisfactory results from the analyses cited above will enable the user to proceed with the intended applications of the model. However, if the performance is found to be unacceptable, the study team will need guidance in rectifying the situation. For example, they might collect additional supplemental field data with which to prepare revised inputs, they might attempt to improve one or more of the technical components of the model, or they might even select a different model and assess its performance. An important consideration that should be addressed in the guidelines is the circumstances (if any) in which the user may proceed with the model applications even though the performance does not meet the standards.

The final important problem we consider here is whether the institutional mechanisms exist that are needed to ensure that the many users of air quality models carry out adequate performance evaluation studies. Problems of particular concern include selecting an organization to oversee model evaluation efforts and determining what services should be provided to the model user community. As the number of model performance evaluations increases, especially those undertaken by relatively inexperienced model users, many problems and questions will undoubtedly arise. There will be a need for a group of air quality modeling specialists to provide the requisite assistance for the users.

In this chapter we have raised some of the more important issues and questions of model performance evaluation. The next chapter discusses these points in more detail.

BLANK PAGE

IV A PROCEDURE FOR EVALUATING MODEL PERFORMANCE

Whenever an air quality model is to be used in a specific application, the questions of the model's suitability and adequacy arise. Therefore, an important part of using an air quality model is an evaluation of that model's ability to provide the desired results. We discussed the general issues involved in model evaluation in the previous chapter; in this chapter we address these issues in more specific terms. This discussion will indicate how the various elements of the model evaluation process might vary depending on the particular type of model and its intended application.

Because proper evaluation of an air quality model involves many individual steps, we begin this chapter with an overview of the process. This overview describes the relationships between the steps, and it identifies the major steps and the subtasks involved. Following the overview, we discuss in detail the entire procedure.

Figure IV-1 presents a simplified flow diagram for the tasks included in a model evaluation. As this figure indicates, we divide the total effort into phases. Before the model evaluation proper is started, the problem requiring the use of an air quality model must be defined. The characteristics of the problem are then examined to guide in choosing a particular model. Model selection is not part of an evaluation study per se, but it does affect the planning of the evaluation study. The complete model evaluation study thus consists of four phases. The first phase is the model selection followed by a series of planning tasks to identify the features of the study. Following these activities is a definition phase, in which all of the study requirements and attributes are formally assembled into a plan for conducting the model performance evaluation. Finally, the evaluation itself is carried out. This fourth phase of the total study includes gathering data, making computer runs, and evaluating the results.

18

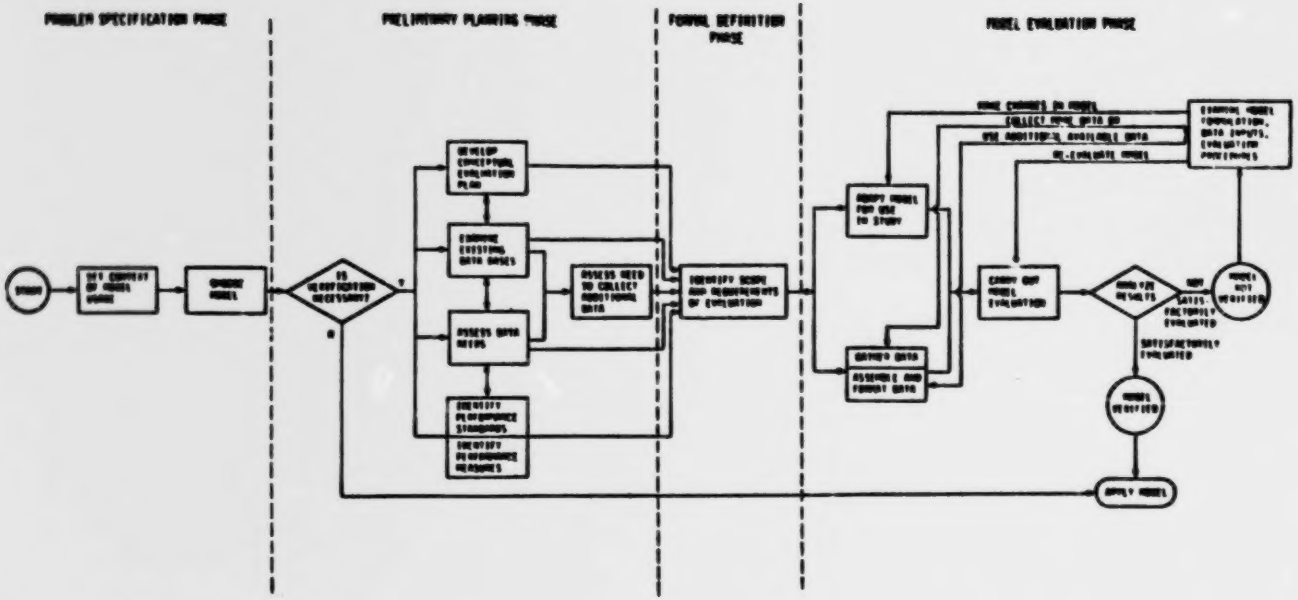

PROBLEM SPECIFICATION PHASE
PRELIMINARY PLANNING PHASE
FORMAL DEFINITION PHASE
MODEL EVALUATION PHASE
START
SET CONTEXT OF MODEL USAGE
SELECT MODEL
IS VERIFICATION NECESSARY?
Y
N
DEVELOP CONCEPTUAL EVALUATION PLAN
EXAMINE EXISTING DATA BASES
ASSESS DATA NEEDS
IDENTIFY PERFORMANCE STANDARDS
IDENTIFY PERFORMANCE MEASURES
ASSESS NEED TO COLLECT ADDITIONAL DATA
IDENTIFY SCOPE AND REQUIREMENTS OF EVALUATION
MAKE CHANGES IN MODEL
COLLECT MORE DATA OR
USE ADDITIONALLY AVAILABLE DATA
RE-EVALUATE MODEL
EXAMINE MODEL FORMULATION, DATA INPUTS, EVALUATION PROCEDURES
ADAPT MODEL FOR USE IN STUDY
GATHER DATA
ASSEMBLE AND FORMAT DATA
CARRY OUT MODEL EVALUATION
ANALYZE RESULTS
NOT SATISFACTORILY EVALUATED
MODEL NOT VERIFIED
SATISFACTORILY EVALUATED
MODEL VERIFIED
APPLY MODEL

FIGURE IV-1. SIMPLIFIED FLOW DIAGRAM OF TASKS IN THE MODEL EVALUATION PROCESS

The elements of a model evaluation, then, can be divided into the following four tasks:

- > Specify the problem and choose a model (not a part of the evaluation study proper).
- > Plan the model evaluation study.
- > Identify the scope and requirements of model evaluation.
- > Perform the model evaluation (i.e., execute the model and assess the results).

Each of those four tasks can be further divided into subtasks. Problem specification consists of subtasks that are primarily involved with assessment of the problem posed and selection of a model to aid in its solution. These subtasks are:

- > Set the context of model usage
- > Choose a model.

Further information on this topic is contained in the EPA (1978a) report.

Once the problem has been defined and the model to be used has been chosen, one must determine if a formal model evaluation is necessary. If it is, the elements of the evaluation study must be identified, the data requirements defined, and appropriate model performance standards and measures chosen. Much of that work can proceed concurrently. For instance, existing data can be gathered at the same time that the complete data requirements are being assessed. Of course, the need for additional data collection can be determined only after data requirements and availability are known. The necessary subtasks in this phase of the evaluation study are:

- > Decide if model evaluation is necessary
- > Develop a conceptual evaluation plan
- > Examine existing data bases
- > Assess data needs
- > Assess the need to collect more data
- > Specify performance standards and measures.

When these subtasks have been completed, the scope and requirements of the evaluation study should be formally laid out. In that third task no new information is developed, but all of the considerations identified earlier are assembled into a plan for the evaluation. The plan is then put into effect. Required data are gathered, the model is adapted for usage in the study area, and pollutant concentrations are predicted and evaluated in light of air quality data and performance standards. The subtasks included in this final task are:

> Modify the model to represent emissions and atmospheric phenomena adequately in the study area.
> Gather required data.
> Assemble and format the data.
> Exercise the model.
> Analyze the results.
> Assess the need to perform further evaluation studies.
> Evaluate the adequacy of the model.

Again, some of these activities may proceed concurrently; model adaptation can be carried out at the same time that data are being collected, for example.

We now proceed to discuss each of these topics individually.

A. PROBLEM SPECIFICATION AND MODEL SELECTION

1. Setting the Context of Model Usage

Setting the context consists of defining the basic problem that necessitates the use of an air quality model. An air quality model is generally used when there is a requirement for more detailed or extensive knowledge than can be obtained from direct analysis of air quality data. Alternatively, a model may be employed to generate information about possible ground-level concentrations of pollutants due to emissions from a source not yet built. In the latter case, an alternative to mathematical modeling* might be to find a similar facility in like surroundings for which an adequate air monitoring data base exists. A study team would be quite fortunate to identify such a situation.

---

* This discussion is limited to mathematical models. Physical models, such as wind tunnels and towing tanks, have yet to be employed in routine air pollution applications.

The first requirement in defining the context of model usage is to delineate the issues to be addressed by the use of the model. Such issues as the formulation and approval of State Implementation Plans require the use of models that can simulate impacts from many sources spread over a wide area. Calculation of impacts from single sources or new projects can be required to satisfy Prevention of Significant Deterioration regulations, New Source Review, and Offset Rules. The preparation of Environmental Impact Statements or evidence for use in litigation can require the use of different types of models, depending on the specific configuration of sources and receptors under review. For a detailed discussion of the various types and applications of air quality models, see the companion report (Hayes, 1979). However, we point out that a primary reason for using an air quality model is the need to quantify air quality impacts of various emissions patterns in order to assess the results of possible scenarios.

When the issues requiring model application have been defined, the characteristics of the situation that guide model selection must be laid out. Such considerations as the pollutants of interest, the applicable air quality standards, the spatial and temporal range of the problem, and the important physical and chemical atmospheric processes to be modeled need to be considered. The pollutants of concern and the proposed model determine the complete set of pollutants that must be modeled. Inert pollutants are regulated in the same chemical form as they are emitted, and thus no chemical reaction processes need be included in the model. Reactive pollutants, by contrast, are subject to chemical transformation in the atmosphere, which the model must take into account. Thus, if O_3 is to be modeled, it is necessary to model its precursors, namely organic compounds, NO, and NO_2. The applicable air quality standard obviously depends on the pollutant considered. Because averaging times in air quality standards for different pollutants range from one hour to one year, the temporal range required of the model is to a large extent determined by the pollutant being modeled. The spatial range of the model application is essentially the domain of "significant influence" of all emissions sources in the study area. These ranges are dependent on meteorological and terrain features of the area as well as on the type of pollutant. The concentrations of primary (inert) pollutants are generally greatest close to their source, and so a large modeling region would be required only if many widely distributed sources were being modeled or if the contaminants were likely to be transported over a relatively large distance. Secondary pollutants (formed by chemical reactions in the atmosphere) reach peak concentrations at some distance from a source and often necessitate the use of a large modeling region. Processes that may need to be treated in a model include pollutant emissions, transport and dispersion, chemical transformations, and removal processes such as rainout, washout, and surface uptake. The relative importance of these processes and the requisite degree of sophistication depend on the specifics of the model application.

2. Selection of a Model

In principle, the air quality model chosen is the one that shows the best match of capabilities to the characteristics of the problem as defined above. Other considerations can also affect the choice of a model. Previous experience with a model and the resulting familiarity with exercising it, for example, will usually result in its being chosen over another comparable model with which the user has little or no experience, provided that both models have similar capabilities and accuracy in performance. Computer access is an obvious prerequisite to the use of some models, and the cost of computer runs to exercise a model can also be a factor in model choice. We note that the EPA (1978a) document provides some guidance in model selection. Efforts to publish comprehensive guidelines that will enable the identification of appropriate air quality models for use in particular applications should continue.

When the problem has been fully defined and an appropriate air quality model has been selected, evaluation of the model in light of the intended use follows. The details of the procedures to be followed in evaluating models depend on both the model type and the application. These specific points are addressed below.

B. PLANNING THE MODEL EVALUATION STUDY

1. Determination of the Need for Model Evaluation

Before a model evaluation is planned, the necessity of such an evaluation should be assessed according to the principle that the user must provide adequate assurance that model performance is satisfactory. In general, we suggest that all modeling efforts include a performance evaluation study. Of course, the extent of each study would depend on the specific model and its intended application. Situations requiring minimal evaluative efforts might occur when:

- > The model is of a type whose performance cannot be evaluated.
- > The application is a New Source Review of a completely new facility where the source has not yet been constructed.
- > The model's performance can be adequately characterized by past experience.

Considering the first case, because of features inherent in model formulation, the performance characteristics of some models cannot be readily determined by the user. Examples of such models include rollback (as applied to photochemical oxidants) and the EKMA. Such models are not verifiable in the sense addressed in this report, because their predictions, as noted in Section E, cannot be checked. Preliminary studies are under way to develop means for assessing the performance characteristics of such models; it would seem that ultimately they will be subjected to a general evalaution. One of the results of the evaluation would be the identification of the input data needed to obtain satisfactory performance. The model user would then be responsible for assembling input data appropriate for the intended study area.

With regard to model formulation, there are two basic ways in which atmospheric processes are treated: (1) as dynamic processes and (2) by invoking a steady-state assumption. In the first case, no problems arise in matching data to model computations because the model formulation is conceptually compatible with the dynamic nature of the atmosphere. In the second case, which includes all Gaussian models, the steady state assumption results in a basic inconsistency between the model and the atmosphere, and the conditions under which data are gathered for an evaluation study might not match those for which the model is appropriate. This problem should be noted in an evaluation study.

Moreover, Gaussian models that compute long-term (annual) average pollutant concentrations invoke a further assumption, namely, that it is appropriate to average over a long period the concentrations calculated using the steady-state assumption. The validity of this assumption needs some attention when evaluating such models. Data collection requirements for long-term models are addressed in a later section.

· In the case of a model application to a source not yet constructed, a full evaluation of the predicted impacts is obviously not possible at that site. At best, the study team could carry out limited tracer experiments. Thus, before a model is used in such an application, it should be verified in a more general setting, e.g., for an existing source of a similar type under comparable meteorological conditions and topographic influences. In this way some confidence could be gained in the use of the model. In addition, such an evaluation would indicate what input data are necessary to permit the model to produce acceptable results. These input data requirements would then need to be satisfied in the new application of the model.

The third circumstance in which model evaluation is not necessary is when the model's performance can be ascertained on the basis of past experience. For example, a model may have been previously verified for identical circumstances. Such a situation could arise when a model, previously used to evaluate the impact of a new highway through an urban area, is employed to evaluate the results of imposing automotive emission control measures in the same area. Another example would be a requirement to evaluate a new emissions control strategy with a model previously used to evaluate a different strategy. In both of these cases the pollutants modeled and the area of application would be identical, and thus the study team could cite the pertinent evaluation effort in lieu of performing further verification work.

Care must be exercised when evaluating the transferability of a previous evaluation. We suggest that criteria be developed to judge applicability. At a minimum, it appears that all of the following criteria would be applicable:

- Similarity of model application. A previous evaluation for the same area is clearly transferable to a new study, but whether a model application in another study area can rely on previous evaluative work is unclear. Ideally, the model should be evaluated each time it is used in a new area. However, with further modeling experience, it may be possible to identify a set of model input requirements that, if satisfied, would ensure adequate model performance. Thus, the user would be responsible for collecting the requisite input data but would not have to mount, say, an additional air monitoring effort for model evaluation purposes.
- Similarity of pollutants modeled. In general, as long as the pollutants are the same, it would appear that the evaluation is transferable. If a model is found to perform satisfactorily for one set of pollutants in some area and if there is an interest in using it to simulate some other set of pollutants, we would recommend that the model be evaluated for the new pollutants.
- Similarity of emissions and meteorology. Even if a model was previously verified for a particular area, significant alterations to the gasoline emissions inventory might necessitate reevaluation of the model. In addition, a verification based on one set of meteorological conditions (e.g., adverse summer episodic conditions) might not be sufficient to ensure adequate performance under a different set of conditions (adverse winter conditions). Again, criteria could be developed to guide the user in judging whether further evaluation of the model is necessary.

In summary, whenever the use of an air quality model is contemplated, it should be a verified model, except for a few restricted cases. The use of an unverified model for any situation involving air quality regulations should be subject to prior approval by the group responsible for overseeing model applications. The circumstances under which the requirement for model evaluation can be waived should be a subject for consideration. It is our view that such waivers should be granted only when exceptional circumstances prevail.

Once the decision has been made to carry out a model evaluation, with the object of verifying an air quality model for a particular use, the process of planning the study can move forward. We now discuss each of the subtasks identified earlier as belonging to the planning phase. Note that they are not necessarily intended to be carried out sequentially, but rather concurrently, as convenience and necessary interaction between them dictate.

2. Development of a Conceptual Evaluation Plan

This initial planning work can be divided into three parts:

> Defining the extent of evaluation required.
> Outlining the implementation of the evaluation study.
> Considering actions to take at the conclusion of the study.

The first part consists of establishing the characteristics and scope of the proposed study; the second part considers the implementation of the effort in light of the available resources; and the third part essentially comprises a contingency plan to anticipate all possible outcomes of the study and what will be done as a result of each. Definition of the extent of evaluation requires that the following issues be addressed:

> Assessment of model availability. When the study is initiated, it should be certain that the latest version of the model is available for use. The potential user should assure himself, by checking with the model developer

or the EPA if necessary, that all known errors in his version of the model have been rectified. If the model is implemented on a computer, a related question that is relevant at this stage is whether the codes are compatible with the user's computer. The magnitude of this problem, particularly for large computer models with extensive file manipulations and overlays, should not be underestimated. Consultation with the model developer may be necessary to effect a transfer.

> Definition of the size and boundaries of the modeling region. The appropriate modeling region is a function of the model type and application. For a Gaussian model applied to a single source, the study has to cover the region where impacts from the source are expected for the meteorological conditions to be modeled. For a multiple source Gaussian climatological model, the modeling region usually covers a complete urban area. With trajectory models, the simulation is constrained by the distances over which the basic model concept is valid. In the case of grid models, a constraint that often comes into play is the number of grid cells that can be accommodated in the available computer core storage. This number of cells has to be judiciously arranged so as to cover the area of interest, or the region segmented, so that only part is dealt with at one time.

> Definition of the resolution of the model. For Gaussian models, which assume steady-state conditions and yield analytical solutions for concentration fields, temporal and spatial resolution is not an issue. (Because of the steady-state assumption the models yield time-invariant concentrations, and because the solution is analytic in spatial coordinates a concentration can be calculated for any desired location.) Other types of models may have particular temporal and spatial characteristics. It should be verified that the temporal resolution of the model is compatible with the

averaging time of the air quality standard being addressed. For short-term computer simulation models, compatibility is not usually a problem with averaging times from 1 hour to 24 hours since time steps in these models are less than 1 hour. However, these models have not been applied to problems involving pollutants subject to annual standards because of the inordinate amount of computing required to simulate such a period. The spatial resolution of the model should be appropriate to the problem being considered. For instance, if the model is being used to establish broad regional trends, a relatively coarsely resolved result might suffice. Studies of mixed control strategies, however, would require that the model be capable of distinguishing the effects of controlling, say, automobiles, power plants, and oil refineries. Control strategy evaluation would necessitate the ability to include the important spatial and temporal characteristics of the emissions patterns associated with these sources, as well as the spatial and temporal characteristics of their respective air quality impacts.

In conclusion, we reiterate that the useful resolution of the model predictions is limited by the characteristics of the available input data. Model results cannot be resolved to a scale finer than that of the input data. For instance, if the diurnal variation in traffic emissions is not properly accounted for, no useful conclusions can be drawn about control strategies that attempt to reduce only peak-hour traffic. If large scale synoptic flows are used as input, no information on local effects of specific terrain features will be available.

> Pollutants to be included. In the case of primary, nonreactive pollutants, only those particular species under study need concern the model user. However, where reactive pollutants are concerned, all species that participate in relevant chemical reactions must be treated. For instance,

to operate the SAI Airshed Model to calculate O_3 concentrations, the user must supply gridded emissions data for hydrocarbons, NO, NO_2, and CO. All of these pollutants affect the predicted O_3 concentrations through reactions in the model's kinetic mechanism.

> Time period to be simulated. This aspect of model operation is the most dependent on the intended application. If compliance for a pollutant subject to an annual average standard is the issue, then a model that can incorporate a year's data will be used. These data are often used in summary form, as an annual stability/wind rose, annual average temperature, and so on. For shorter term studies, such as 1-hour-average concentrations, the period simulated generally varies from 1 hour up to 24 hours or possibly longer. With a model that does not include any temporal variation, only a steady-state concentration is calculated. For photochemical models that include temporally varying inputs, the period generally simulated is the daylight hours, when photochemical reactions are important. Thus, the variations of emissions and meteorological conditions over this time must be included. Sometimes, pollutant carry-over from the previous day is important and must be treated through the performance of a multiple-day simulation.

> Listing of the kinds of data that will be needed. As a preliminary task in determining the scope of the required evaluation, the types of data required should be identified. At this stage, detailed data requirements are not listed, but knowledge of the types of data required will enable a proper assessment of current availability of useful data and the probability that a suitable data collection effort can be initiated. Table IV-1 presents a list of types of data that may be required by an air quality model. The level of detail required by any individual model is specific to that model.

TABLE IV-1. POSSIBLE DATA REUQIREMENTS OF AIR QUALITY MODELS

Data Category	Examples
Meteorological	Wind speed and direction
	Temperature
	Atmospheric stability
	Mixing depth
	Insolation
	Humidity
	Cloud cover
	Atmospheric pressure
Emissions	Mobile sources
	Stationary sources
	Natural sources
Air quality	Surface measurements
	Vertical pollutant concentration soundings

For instance, a single source model usually requires wind speed and direction inputs only at the point of emissions. For a long-term average model, a wind rose covering the appropriate length of time is needed. A grid-based photochemical model requires a three-dimensional, temporally varying wind field. Thus it is useful to list the types of data needed for the model so that the extent of data gathering required can be appreciated.

> Choice of meteorological conditions for study. Since meteorological conditions play a significant role in determining pollutant concentrations, the choice of conditions for a modeling study is very important. Generally, episodic conditions leading to high pollutant concentrations are desirable. These can be identified, for locations that have the necessary data, by

examining the concentration records for a period in the past and observing which meteorological conditions lead to the highest observed concentrations. Of course, this approach works only if there are emissions sources in the area: It is not applicable to a hypothetical emissions source in an area with clean air. If air quality records are not available, an analysis of the meteorological data must be carried out to identify adverse conditions. Conditions judged most likely to lead to significant impacts should be identified (e.g., low mixing depth and stagnant wind), and their historical frequencies of occurrence should be determined. For a short-term simulation, data for one or more days in the past will need to be collected and used in the model evaluation effort.

> Determination of the number of meteorological regimes to study. The basic considerations in selecting the number of regimes to be included in an evaluation study are (1) specification of a sufficient number to enable characterization of model performance, and (2) selection of regimes representative of those required in the actual applications studies. While little definitive guidance can be provided at the present time, we do offer the following observations. With regard to long-term average models, Rubin (1974) found that year-to-year variations in annual stability/wind rose had a significant effect on pollutant concentrations calculated by the Air Quality Display Model. He concluded that such year-to-year variations would have an important effect on model validation if a rose were used from a year other than that for which pollutant concentrations were recorded. Thus, roses from several years should be used so that variations are accounted for. For model evaluation, pollutant concentrations for several years would have to be available, and normally they are not.

Ideally, for short-term average models, all frequently occurring adverse meteorological regimes would be examined. However,

> it should be necessary for the user to evaluate the model subject only to those regimes that will actually be employed in subsequent applications. For example, if a model is to be used to estimate O_3 concentrations that occur under adverse conditions during the summer, it would not be necessary to evaluate the model for conditions that occur during the winter. We note that photochemical model usage generally entails the selection of conditions leading to high O_3 concentrations. For these models, it would also be advisable to select conditions resulting in lower maximum O_3 levels (e.g., from 0.12 to 0.20 ppm) to test the model's performance at concentrations close to the standard (0.12 ppm). We also note that the number of conditions to be studied must be reconciled with financial resources.

The next step in developing the conceptual evaluation plan is to lay out the resources available to implement the plan. These resources are discussed below.

a. Financial Resources and Temporal Constraints

At this stage in the evaluation study the complete requirements have not been identified, and so the dollar needs of the study cannot be specified accurately. A rough estimate can be made, however, on the basis of the extent of the study (discussed above) and previous similar model evaluation efforts. Funds for the evaluation can come from both public and private sources: Federal, state, and local governmental agencies may fund a model evaluation study as part of an effort like the development of a State Implementation Plan, while private sources might fund evaluation efforts in conjunction with a New Source Review.

The responsibility for financial support of model evaluation efforts is a matter deserving some consideration. In general, funding should be provided for studies aimed at collecting requisite data bases and carrying out general evaluations of EPA-recommended models. Consideration should also be given to setting aside additional funds for model verification, which would be made available to public agencies about to use an air quality model (new or established) in a new application. We discuss this point further in the next chapter.

An issue related to the economic resources available for model evaluation is the amount of time available to complete the study. In many cases the model application must be completed by a definite date. For example, a provision in a law may require certain actions by certain dates, or a commitment must be honored to commence construction of a project by a particular date. It will be important for the study team to carry out the evaluation planning task in a timely fashion to allow adequate time to actually evaluate the model's performance. Moreover, they should be realistic with respect to the total time required for the study.

Any problems associated with time or money limitations must be resolved in order that an adequate plan can be formulated. However, this resolution should take place at the formal plan definition stage, when the plan for the study has been completed. If the amount of work required by the plan is greater than the resources available to carry it out, then it must be explicitly recognized that the study will be inadequate and the impact of the deficiencies on the study's conclusions should be estimated. Priorities should be assigned to the various elements of the plan to ensure that the most important activities are carried out.

As an example, if only one vertical temperature sounding is available for an area, the consequences of having so little information for characterizing mixing depths and winds aloft must be established. For models that do not use detailed spatial and temporal information, the number of temperature soundings may be of little consequence, but for a complex model the number could have a substantial impact on the model's prediction. We note that the effect of data limitations on the adequacy of predictions of photochemical grid models is a subject of current research studies (Tesche, 1978).

b. Manpower Resources

To carry out a model evaluation study, people with many different skills are required to undertake many different types of tasks. Personnel are needed with experience in the fields of:

> Meteorology
> Analytical chemistry
> Computer programming
> Statistical analysis
> Air quality analysis.

These people are needed to:

> Select an appropriate air quality model
> Plan and set up the evaluation study
> Collect data
> Specify and prepare data inputs to the model
> Run the model
> Analyze the study results.

In addition, supervisory personnel are needed to assume responsibility for proper execution of all of these activities.

In view of the likelihood that a relatively large number of model evaluation efforts may be initiated in the future and that there will be a need for personnel with a high level of expertise to design and implement these studies, some form of central model evaluation group should be established. The function of this group would be to be available to help model users who need to set up a model evaluation study. The group would include experts in the fields listed above and would be available to help plan or review studies and aid in their implementation. Thus there would be experts available for advice on model selection and usage, air quality measurements, meteorological measurements, assembly of emissions inventories, setting up and running air quality models on the computer, and analysis of the results. In addition, the group could serve a custodial function in collecting a set of properly evaluated air quality models and a number of data bases to use in future evaluations. The availability of such a group would relieve model users of the responsibility to assemble a comparable group of experts (presumably from outside contractors) and, at the same time, would build up a fund of knowledge of and expertise in model evaluation. Accumulation of this knowledge would also result in improvements in model evaluation methods over a period of time.

c. Hardware Resources

Hardware will be needed for data collection (to the extent that necessary data are unavailable), exercise of the model, and analysis of results. Table IV-2 lists the types of hardware that can be required to collect the data necessary to exercise the model. For running the model and analyzing the results, access to computing facilities is often required. A portion of the measurement hardware may be available as part of a currently operating data-gathering network. All available aerometric monitoring hardware should be inventoried, with special consideration given to the identification of those instruments that can be deployed at the discretion of the model evaluation group. In the early stages of the model evaluation effort, the exact extent of the data-gathering effort will not be known, but a survey of the currently available instrumentation will indicate the size of the effort that can be mounted without purchasing or leasing equipment.

Once the extent of the evaluation effort has been defined and the available resources are known, the final part of the plan for the evaluation study can be undertaken, which is to lay out actions to be taken at the conclusion of the study. The product of the performance evaluation consists of values for one or more performance measures that are to be compared with appropriate standards. In the event that satisfactory performance is achieved, the results of the study can be reported and the model can be considered

TABLE IV-2. POSSIBLE HARDWARE REQUIREMENTS FOR DATA COLLECTION

Data Category	Hardware
Meteorological	Anemometers (wind speed)
	Wind vanes (wind direction)
	Thermometers (temperature)
	Instrumented balloons (upper air measurements of winds and temperatures)
	Radiometers (insolation)
	Hygrometers (relative humidity)
	Barometers (atmospheric pressure)
Emissions	Traffic counters
	Source monitors
Air quality	Samplers
	Analytical instrumentation
	Calibration samples
	Recorders
	Instrumented aircraft

to be verified. If the model fails to achieve the appropriate performance standard, the possible reasons for failure must be examined. These reasons can include:

- > Inadequate data base.
  - Data too sparse.
  - Data too imprecise.
- > Inadequate model inputs.
  - Uncertain interpolation of data.
  - Uncertain extrapolation of data.
- > Inadequate model formulation.
  - Poor treatment of emissions.
  - Inadequate transport and dispersion algorithms.
  - Inadequate chemical mechanism for atmospheric reactions.
  - Poor treatment of removal processes.
- > Inadequate air quality data.
  - Too few data to calculate performance measure properly.
  - Data uncertainties too great.

We discuss these possibilities in detail in a later section. At this stage the implications of an unsuccessful evaluation study should be considered. It would seem appropriate to set aside some time and financial resources for possible use at the end of the study to enable the performance of some additional work to achieve a set of satisfactory model results.

3. Examination of Existing Data Bases

In this task, existing data bases are surveyed with a view to their use for the model evaluation, To the extent that suitable data are already available, they will not need to be collected as part of a special field program. We anticipate that many model evaluations can make significant use of existing data. Data required for the evaluation are of three basic types:

- > Meteorological data
- > Emissions data
- > Air quality data.

They may be available from three sources:

- > Federal government agencies
- > State and local government agencies
- > Private organizations.

Data from governmental sources should be readily available. Data from the private sector may be harder to obtain and, in addition, may be of limited utility, since they may have been collected for some special purpose.

The characteristics of the data bases to be evaluated are:

- > Coverage of the modeling region.
- > Coverage of the desired range of meteorological conditions.
- > Coverage of the required time period.
- > Coverage of the required model variables.
- > Compatibility with the model's spatial and temporal resolution.

These characteristics are all related to the features of the study that were identified in developing the conceptual plan for the evaluation.

Coverage of the modeling region is judged by the adequacy with which the data base collection system covers all data of interest. For a study involving a single point source, this area would include the location of the source's maximum impacts under the various meteorological regimes of interest (and the location of the source itself). For a regional scale study, the area of interest might extend beyond the actual region modeled to include outside sources that impact on the region. The boundary of the region of interest is also influenced by terrain features, such as mountain barriers. If pollutants do not traverse such features, there is usually no interest in modeling the opposite side. Also, if pollutants are carried outside the region during one day and brought back by wind the next day, there would be a need to locate monitors to gather the relevant data.

The existing data base should also be judged for its coverage of the desired range of meteorological conditions. In the course of defining the extent of the evaluation study, one should select the meteorological conditions most appropriate in light of the intended model application. Meteorological and air quality data should be assembled for these conditions.

Another feature of the evaluation study that was defined in the planning stage is the period of time over which the data should be collected. For models that simulate a day or less, available data bases are likely to be able to provide data, although adequacy of areal coverage or availability of data for a specific day may be a problem. For situations requiring the estimation of annual average concentrations, care should be taken that the data do not include systematic gaps (e.g., all summer data missing). Seasonal influences on emissions may have to be taken into account in preparing an emissions inventory. Furthermore, for a satisfactory evaluation, a sufficient number of years of data should be available to constitute a representative sample.

The variables for which data are required and the temporal and spatial resolution of those data should also be identified and listed at the planning stage. Existing data bases must be examined for comprehensiveness and compatibility with model requirements. Data requirements for which existing data are not suitable or comprehensive enough must be fulfilled in the course of a supplementary field measurement program. However, only certain types of supplementary data collection are possible. If some measurements on a particular day were missed, it is impossible to collect them subsequently. Additional days of data might, however, be collected if necessary. If existing data are used anyway, the consequences of their use must be recognized. The effect of degradation of model inputs on model output should be determined for each EPA-recommended model through the performance of appropriate sensitivity studies. For a simply formulated model, determination of this sensitivity to data degradation may be straightforward, but for a complex model the work involved will be substantial. Plans for carrying out one such effort for a complex model are described by Tesche (1978).

The largest single source of meteorological data is the National Weather Service, which supervises the collection of data at a network of stations, mostly located at airports, around the country. Other sources of meteorological information are air pollution agencies, which often record surface wind and temperature information. However, routine data collection activities are likely to be sparse, as illustrated by Table IV-3, shows the number and type of daily meteorological measurements made in 15 U.S. cities (Tesche, 1978). The level of monitoring varies widely from city to city, and it will not be possible to say how many data are available for any given area without surveying all possible sources.

The major repositories of emissions data are, of course, the air pollution agencies. The EPA maintains a system, called the Aerometric and Emissions Reporting System (AEROS), that contains both emissions and air quality information reported by pollution control agencies throughout the United States. The emissions data are stored in the National Emissions Data System (NEDS), which has the capability of storing and retrieving source and emissions-related data for particulates, SO_x, NO_x, CO, and hydrocarbons. A problem with this centralized system is that the data can never be fully up to date. To obtain the most recent data, one must consult appropriate local or state authorities. Table IV-4 summarizes the types of emissions inventories available for the 15 cities listed in Table IV-3. Note that the level of detail available can vary from disaggregation on a 0.1 km grid down to aggregation on a county-wide scale. The latter scale would clearly be inadequate for some models, and a supplementary emissions preparation effort would need to be considered. This topic is covered later.

Air quality data are required for the model evaluation not only to supply input to the model, but also to provide the basis for comparison of model results with performance standards. Input air quality data are used to supply background pollutant concentrations for many models and to specify initial and boundary conditions for complex models. The Storage and Retrieval of Aerometric Data (SAROAD) system is a primary source of air

TABLE IV-3. NUMBER AND TYPE OF DAILY METEOROLOGICAL MEASUREMENTS IN 15 U.S. CITIES IN 1977

City	Surface Wind Velocity	Surface Temperature	Atmospheric Stability	Upper Level Wind Velocity	Solar Insolation	Humidity
Albuquerque, NM	7	7	RW_1	RW_1	1	1
Chicago, IL	10	10	RW_2	RW_2	3	3
Denver, CO	25	2	RW_1	RW_1	1	1
Houston, TX	3	3	0	P_1	1	1
Las Vegas, NV	8	1	AC_2, AS_1	0	1	1
Los Angeles, CA	44	9	RD_8, AS_1	RD_8	2	8
New York, NY	10	10		P_1	3	3
Philadelphia, PA	2	2		P_1	1	1
Phoenix, AZ	8	8	0	0	1	1
Portland, OR	9	9	RW_1	RW_1, P_1	1	1
Sacramento, CA	12	4	AS_3	P_4	1	2
San Diego, CA						
San Francisco, CA	17	17	RW_1	RW_1	17	17
St. Louis, MO	25	25	RD_8	P_{96}	6	20
Washington, D.C.	25	25	RW_1	RW_1	2	2

AC = acoustic sounder.

AS = aircraft spiral.

RD = radiosonde.

RW = rawinsonde

P = pibal.

Note: Subscripts refer to the number of measurements taken each day. A zero entry indicates that a particular measurement is not taken; a blank indicates uncertainty as to whether or to what extent the measurement is taken.

Source: Tesche (1978).

TABLE IV-4. METHODS USED FOR THE PREPARATION OF EMISSIONS INVENTORIES IN 15 U.S. CITIES IN 1977

City	Format	Species	Grid Size	Hot/Cold Start	Format	Species	Spatial	Temporal
Albuquerque, NM	Link-node: VMT	N, H, C	N/A	Area-wide temporal resolution	NEDS	N, H, S, P, C	Area-wide	Annual average
Chicago, IL	Gridded	N, H, C	50 x 50: 2 mi		Gridded	N, H, C	2 mi	
Denver, CO	Gridded	N, H, S, P, C	30 x 30: 1 mi	Area-wide temporal distribution	Gridded	N, H, S, P	1 mi	8 or 24 hour, plus seasonal
Houston, TX	Link-node: VMT				By counties	H, S, H, P	County-wide	
Las Vegas, NV	Gridded	N, H, C	30 x 40: 1 km	Area-wide temporal distribution	Gridded	N	1 km	
Los Angeles, CA	Gridded	N, H, S, C	100 x 50: 2 mi	Area-wide temporal distribution	Gridded	N, H, S, C	2 mi	Hourly
New York, NY	VMT		Borough by borough			S, P		
Philadelphia, PA	Gridded	H, C, N	48 x 48: 2 mi		NEDS	N, H, S, P, C	Area-wide	Annual average
Phoenix, AZ	Gridded	N, H, C	1 mi		NEDS	N, H, S, D, C	Area-wide	Annual average
Portland, OR	Gridded	S, P, C	20 x 30: 2 km		By district	H, N	Depends on size of districts	
Sacramento, CA	Gridded	N, H, S, P, C	25 x 25: 2 km	Area-wide temporal resolution	Gridded	N, H, S, P, C	2 km	Annual average
San Diego, CA								
San Francisco, CA	Gridded	N, H, S, P, C	120 x 60: 1 km	Area-wide temporal distribution	Gridded	N, H, S, P, C	1 km	Hourly
St. Louis, MO	Variable size grid	N, H, S, P, C	150 x 200: 1-10 km	Hot/cold distributions applied to each grid cell	Gridded	N, H, S, P, C: hydrocarbon speciation	1-10 km	Hourly
Washington, D.C.	Gridded	N, H, S, P, C	4 mi		Gridded	N, H, S, P, C	4 mi	

N = nitrogen oxides.

H = hydrocarbons.

S = sulfur oxides.

P = particulates.

C = carbon monoxide.

VMT = vehicle miles traveled.

NEDS = National Emissions Data System.

Source: Tesche (1978).

quality data for use in modeling studies. SAROAD is maintained by the EPA as part of AEROS, which is used to store and report air quality and emissions data. The data in SAROAD may not be sufficient, however, because in many areas the reporting network is based on the minimum number of stations required by the EPA (40 CFR §51.17, 1975) (see Exhibit IV-1). Unfortunately, there are currently no criteria for judging the adequacy of these monitoring networks for use in evaluating model performance. In fact, it is likely that the monitoring objective of an existing network is something other than to provide information to the model evaluation process. As a result, locations of the monitors may make those measurements of limited usefulness in an air quality modeling study. In particular, existing monitoring facilities will likely be concentrated in areas where air quality violations occur, which tend to be in the centers of regions. Thus, data that are currently available often do not give information on boundary or background concentrations. We discuss placement of monitors in a later section.

Table IV-5 presents the extent of routine air quality monitoring in the 15 cities listed in Tables IV-3 and IV-4. Again, there is a noticeable variation in the extent of monitoring from city to city, with the amounts generally corresponding to the severity of air pollution problems.

In summary, there are many sources of relevant meteorological, emissions, and air quality data, but unless the data are obtained from a special intensive monitoring activity whose objective is to provide model input information, they may not be comprehensive enough to satisfy the complete needs of a model evaluation effort. A relevant matter not covered in this section is the problem of data quality. We discuss this issue more fully in the section on assessing data needs, but we point out here that data quality is a very important element in judging the acceptability of a data base.

4. Assessment of Data Needs

The data needs for a model evaluation study are dependent on:

- The model.
- The nature of the evaluation (general or site-specific).
- The relationship between amount and quality of data and model performance.

Classification of region	Pollutant	Measurement method or principle [i]	Minimum frequency of sampling	Region population	Minimum number of air quality monitoring sites [h]
I	Suspended particulates	High volume sampler	One 24-hour sample every 6 days [a]	Less than 100,000	4.
				100,000–1,000,000	4+0.6 per 100,000 population.[c]
				1,000,001–5,000,000	7.5+0.25 per 100,000 population.[c]
				Above 5,000,000	12+0.16 per 100,000 population.
		Tape sampler	One sample every 2 hours		One per 250,000 population [c] up to eight sites.
	Sulfur dioxide	Pararosaniline or equivalent [d]	One 24-hour sample every 6 days (gas bubbler).[a]	Less than 100,000	3.
				100,000–1,000,000	2.5+0.5 per 100,000 population.[c]
				1,000,001–5,000,000	6+0.15 per 100,000 population.[c]
				Above 5,000,000	11+0.05 per 100,000 population.
			Continuous	Less than 100,000	1.
				100,000–5,000,000	1+0.15 per 100,000 population.[c]
				Above 5,000,000	6+0.05 per 100,000 population.[c]
	Carbon monoxide	Nondispersive infrared or equivalent.[e]	Continuous	Less than 100,000	1.
				100,000–5,000,000	1+0.15 per 100,000 population.[c]
				Above 5,000,000	6+0.05 per 100,000 population.[c]
	Photochemical oxidants	Gas phase chemiluminescence or equivalent.[f]	Continuous	Less than 100,000	1.
				100,000–5,000,000	1+0.15 per 100,000 population.[c]
				Above 5,000,000	6+0.05 per 100,000 population.[c]
	Nitrogen dioxide	24-hour sampling method (Jacobs-Hochheiser method).	One 24-hour sample every 14 days (gas bubbler).[b]	Less than 100,000	3.
				100,000–1,000,000	4+0.6 per 100,000 population.[c]
				Above 1,000,000	10.
II	Suspended particulates	High volume sampler	One 24-hour sample every 6 days [a]		3.
		Tape sampler	One sample every 2 hours		1.
	Sulfur dioxide	Pararosaniline or equivalent [d]	One 24-hour sample every 6 days (gas bubbler).[a]		3.
			Continuous		1.
III [g]	Suspended particulates	High volume sampler	One 24-hour sample every 6 days [a]		1.
	Sulfur dioxide	Pararosaniline or equivalent [d]	One 24-hour sample every 6 days (gas bubbler).[a]		1.

[a] Equivalent to 61 random samples per year.

[b] Equivalent to 26 random samples per year.

[c] Total population of a region. When required number of samplers includes a fraction, round-off to nearest whole number.

[d-f] [Reserved]

[g] It is assumed that the Federal motor vehicle emission standards will achieve and maintain the national standards for carbon monoxide, nitrogen dioxide, and photochemical oxidants; therefore, no monitoring sites are required for these pollutants.

[h] In interstate regions, the number of sites required should be prorated to each State on a population basis.

[i] Named methods and principles, except the tape sampler method, are described in Part 50 of this chapter. The tape sampler method is described in Hemeon, W. C. L., Haines, G. F., Jr. and Ide, H. M., "Determination of Haze and Smoke Concentrations by Filter Paper Samplers", J. Air Pollution Control Association, Vol. 3, pp. 22-28, 1953. Use of these and other methods shall be as specified in § 51.17a.

Source: 40 CFR §51.17 (1975).

EXHIBIT IV-1. AIR QUALITY MONITORING REQUIRED BY THE EPA FOR VARIOUS POLLUTANTS

MICROFILMED FROM
BEST AVAILABLE COPY

TABLE IV-5. NUMBER OF STATIONS PERFORMING ROUTINE AIR QUALITY SAMPLING IN 15 MAJOR CITIES IN THE UNITED STATES

City	Oxidant	NO_x	CO	SO_2	RHC	Particulates	$SO_4^=$	Upper Air Measurements	Hydrocarbon Species
Albuquerque, NM	4	3	5	3	0	13	0	0	0
Chicago, IL	4								
Denver, CO	9	3	9	1	2	0	0	R	
Houston, TX	3	3	3	3	3	3			
Las Vegas, NV	3	2	24	0	4	3	0	0	0
Los Angeles, CA	39	27	23	17	11		8	R	S
New York, NY	7	7							
Philadelphia, PA	8	5	8	8	3	8			
Phoenix, AZ	5	2	8		3				
Portland, OR	3	3	15						
Sacramento, CA	8	8	8		4	1	0	0	
San Diego, CA								0	
San Francisco, CA	26	16	16	21	16	17	17	0	
St. Louis, MO	25	25	25	11	25	20	10	S	S
Washington, D.C.	10	10	10	10	10	10			

S = special studies.

R = rarely.

Note: A zero entry indicates that a particular measurement is not taken; a blank indicates uncertainty as to whether or to what extent the measurement is taken.

Source: Tesche (1978).

The assessment of what data need be collected should be carried out concurrently with the examination of existing data bases. These two tasks should be executed with a high degree of interaction. The need for particular data should initiate a search for those data among existing data bases. Conversely, the total lack of a particular type of data and the lack of any prospect of obtaining them could modify the assessment of data needs. For instance, lack of vertical temperature soundings at a sufficient number of locations might be compensated for by using a single temperature sounding in conjunction with data from several surface temperature observation sites to prescribe the spatial and temporal characteristics of the mixing depth.

The question of the number and location of monitoring sites required for air quality model evaluation is at present unresolved. The number of stations employed in evaluative studies at the present time is generally determined by the available resources. Obviously, there is a need for an effort to determine more accurately how many measurements are needed to satisfy the various requirements of the model evaluation. The number will, of course, depend on the model and the application.

There is also a need for an effort to develop an optimal siting methodology for pollutant monitoring stations, despite many recent studies dealing with this subject. Seinfeld (1972), for example, developed a siting algorithm based on the premise that stations should be located so that the concentration data are as sensitive as possible to changes in emissions from major sources. Hougland and Stevens (1976) developed a site location model based on maximizing the sum of coverage factors for each source. Ott's (1977) standardized system of site selection is designed to improve the comparability of data from different stations. Pooler (1974) described the rationale behind the selection of sites for the St. Louis RAPS study and its relation to monitoring objectives. Liu et al. (1977) developed a methodology for designing monitoring networks based on a figure of merit that was related to the probability of detection of concentration peaks. Although there have been many studies of network designs, using many different algorithms, a practical methodology suitable for siting stations for model evaluation purposes has yet to be developed.

As mentioned earlier, all models require data of three basic types:

- Meteorological data
- Emission data
- Air quality data.

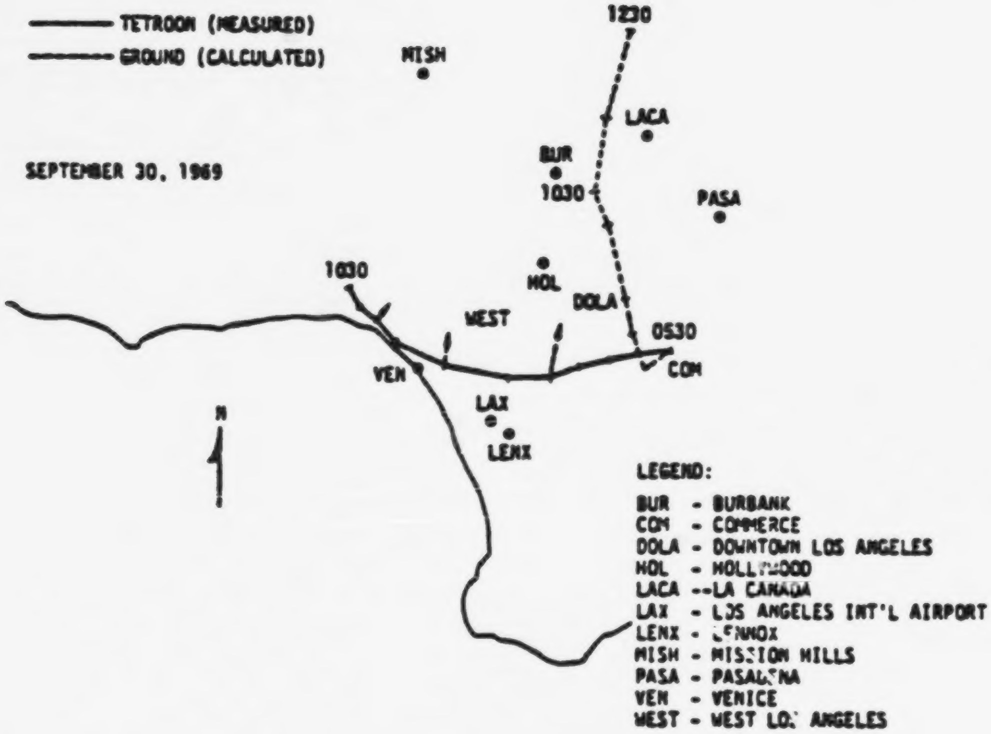
In general, the more complex a model is, the more flexibility it allows in the preparation of the input data. We now discuss the data requirements of different models for each of the three categories listed above. Our discussion is general, with specific comments made about individual model types as appropriate. We then discuss data quality control.

a. Meteorological Data

One of the principal mechanisms for distribution of pollutants is transport by the wind. Therefore, wind speeds are required inputs for many air quality models. In the simplest case, that of a short term Gaussian model, the wind speed at the emissions source is required. Since such a model is commonly used to compute concentrations resulting from emissions from elevated stacks, the wind speed aloft is the relevant parameter. However, the vast majority of wind speed measurements are made at or near ground level. Extrapolation of surface measurements to upper levels is commonly done by using an empirical power law correlation, which can introduce errors on the order of 20 percent at 2000 feet (Roth et al., 1975). In the case of Gaussian models that simulate longer term (up to annual) average concentrations, winds are needed over correspondingly longer periods. They are supplied as a stability/wind rose, which is a joint frequency function for wind speed, wind direction, and atmospheric stability. Since a single rose is used for the complete modeling area, the questions arise of where the measurements are made and how typical they are. The importance of wind speed in the Gaussian formula argues for its careful measurement in a model evaluation study. Enough measurements of wind speeds at ground level and aloft should be made to check any correlation used for extrapolation. In addition, measurements of wind roses at a number of locations in the modeling area should be made to ascertain the potential variations that can occur and to estimate the potential errors from this source. The exact number of locations needed will have to be determined from experience gained in a few studies. If it is not possible to make wind rose measurements for the needed length of time (up to a year), then they should be made for as long as possible to obtain at least a lower bound on the potential error.

For more complex models, such as grid and trajectory models, more detail yet is required in the wind measurements. Trajectory models, which track a moving parcel of air and neglect the variation of wind speed with height, must have some means of converting measurements into a characteristic wind speed and direction for the parcel as a function of time. Use of ground-based measurements can lead to large errors in the computed trajectory, as illustrated in Figure IV-2. This potential for error suggests that an element of evaluating a trajectory model might include the comparison of computed trajectories with suitable observations, such as tetroon releases, and an accounting for any discrepancies found. Another type of trajectory model is the Reactive Plume Model (RPM) (Liu, Stewart, and Roth, 1978). In this model no spatial variation of the wind is assumed, just a temporal variation. However, the winds should be defined at the height of the source, and so the previous remarks about wind speed variations with height apply.

Grid models utilize a wind field defined hourly on a two- or three-dimensional grid over the time period of the simulation. Definition of such a wind field requires both surface and upper air observations. When wind data aloft are not available, theoretical wind shear relationships must be employed. As noted above, no definitive guidance is available as to the appropriate number and location of monitors. For general guidance, Hayes, Reynolds, and Roth (1977) recommended a minimum of 8 to 12 monitors and a desirable number of 12 to 20, making continuous measurements, for characterizing ground-level wind speeds and directions. For vertical wind soundings, they recommended a minimum of one location measuring four times per day and a desirable number of three locations making measurements six to


TETROON (MEASURED)
GROUND (CALCULATED)
SEPTEMBER 30, 1969
1230
MISH
LACA
BUR
1030
PASA
1030
HOL
DOLA
WEST
0530
COM
VEN
LAX
LENX
N
LEGEND:
BUR - BURBANK
COM - COMMERCE
DOLA - DOWNTOWN LOS ANGELES
HOL - HOLLYWOOD
LACA - LA CANADA
LAX - LOS ANGELES INT'L AIRPORT
LENX - LENNOX
MISH - MISSION HILLS
PASA - PASADENA
VEN - VENICE
WEST - WEST LOS ANGELES

Source: Eschenroeder, Martinez, and Nordsieck (1972).

FIGURE IV-2. COMPARISON OF CALCULATED GROUND TRAJECTORY WITH OBSERVED TETROON TRAJECTORY FOR THE LOS ANGELES BASIN

eight times per day. These recommendations were for a verification study in a hypothetical urban area with a population of 1 million people and an area of about 2500 km^2. The features of a particular area that may affect the number of monitors required are:

> Topography. Where there is significant influence on the wind flow by terrain features, extra monitors may be required near those features to define the flow fields adequately.
> Emissions patterns. Where there is high spatial variability in the pollutant concentration pattern, we can expect comparable variability in the pollutant concentration patterns. As a result, careful definition of the wind and air quality monitoring locations will be required. Where emissions are evenly spread over a region, concentration patterns will be smoother and less sensitive to small fluctuations in wind speeds and directions.
> City size. Cities covering large areas should have more monitors to treat adequately possible complexities in the pollutant concentration field.

Other meteorological data required for large, computer-based air quality simulation models are temperature soundings (used to obtain mixing depth and atmospheric stability information), insolation (solar radiation intensity), and cloud cover. Much the same considerations apply to these variables as to wind measurements, but in general fewer measurements are required to define their input values adequately. Table IV-6 shows the meteorological data requirements recommended by Hayes, Reynolds, and Roth (1977). Satisfactory mixing depth data are important because computed concentrations are often nearly proportional to this variable. Measurements of insolation and cloud cover are required when photochemically reactive pollutants are to be modeled or when they are needed to estimate atmospheric mixing characteristics.

TABLE IV-6. ESTIMATED METEOROLOGICAL DATA REQUIREMENTS FOR EVALUATION OF A LARGE AIR QUALITY SIMULATION MODEL IN A HYPOTHETICAL URBAN AREA

Parameter	Number of Stations		Frequency of Measurements (observations per day)	
	Minimal	Desirable	Minimal	Desirable
Wind speed and direction (ground-based)	8-12	12-20	Continuous*	Continuous*
Vertical wind soundings	1	3	4	6-8
Temperature soundings	2	4	4	6-8
Insolation	1	3	Continuous*	Continuous*
Cloud cover	1	3	8	24

---

* Continuous indicates hourly averaged data.

Source: Hayes, Reynolds, and Roth (1977).

b. Emissions Data

To calculate atmospheric pollutant concentrations, one must first have full information on the quantities of pollutants and precursors being emitted. Such information is obtained in the form of an emissions inventory. A complete inventory has data on emissions from stationary sources (including large point sources, such as power plants, and distributed area sources such as home heating furnaces), mobile sources (principally automobiles, trucks, and aircraft), and natural sources (such as vegetation).

The extent and degree of detail needed for an emissions inventory depends on model requirements and the application. Data on the emissions of a single point source should be readily obtainable. However, the assembly of a detailed emissions inventory for a number of pollutants over a large urban area can be an expensive and time-consuming task, involving as much if not more effort than the modeling exercise itself; general guidelines for preparing an emissions inventory are given in the reports by EPA (1973). The ease with which an emissions inventory can be assembled depends on what kind of inventory is already available. If it is extensive as the result of ongoing surveillance activities, relatively little extra effort may be needed, but if little information is available, a trade-off will need to be made between the effort involved to assemble a full inventory and the degradation of results associated with an imcomplete inventory.

When data on emissions from a single large point source or a small group of sources are required, they can be obtained directly through stack measurements or indirectly from knowledge of the throughput or operating characteristics of the specific equipment. If a large number of plants are being modeled or specific information is not available, emissions can be calculated for general classes of sources through the use of the EPA's "Compilation of Air Pollutant Emission Factors" (EPA, 1972), which gives estimates of emissions for many types of equipment as a function of some activity level such as fuel consumption. These factors are averages estimated with varying degrees of accuracy; in any case they would not be expected to be as accurate as specific information on a particular piece of equipment.

In construction of emissions inventories, small point sources are generally aggregated and treated as area sources because individually each one is too small to have a significant impact, but a large number of such sources spread over a few square kilometers will contribute an appreciable amount of pollutants. In some applications, where short-term-average concentrations are required, the temporal variations in emissions from these sources are important (for instance, residential space heating has pronounced seasonal and diurnal variations). If an annual average concentration is sought, only larger-scale temporal fluctuations (e.g., seasonal) may be required for the model inputs.

Mobile source emissions are obtained by combining traffic and vehicle driving chracteristics with emissions rates for individual types of vehicles (EPA, 1978b). Different levels of detail are possible, ranging from an overall estimate of vehicle miles traveled (VMT) combined with characteristics of an assumed vehicle population, to detailed characterization of traffic density and speeds for all major streets by time of day, together with emissions data from a representative sample of the actual vehicle population. Traffic and vehicle data are often available from local highway departments that collected them while planning transportation system improvements.

A relatively low level of detail in mobile source emissions inventory would be required if the contribution of traffic to annual average particulate concentrations is to be studied. In some models, merely an annual average, not diurnal and seasonal variations in traffic densities would be required. By contrast, if the formation of photochemical pollutants is to be followed over a wide area on an hourly basis, an emissions inventory with substantial spatial and temporal detail is required. In the absence of specific information, many assumptions must be invoked, with the concomitant possibility of degradation in the quality of model results. For example, vehicular emissions are dependent on vehicle operating characteristics. Therefore, to account properly for photochemical precursors emitted during the morning rush hour, the traffic distribution and average speed patterns on surface streets and freeways during those hours are needed. Note that collection of such detailed data where they are not previously available can be a time-consuming and expensive task.

As pointed out above, any model application requires full information on emissions of all pollutants and precursors of interest. This specification may be relatively simple in the case of a model used for one inert pollutant, but for a complex model incorporating a comprehensive photochemical reaction mechanism, complete data may not be available. For instance, the SAI Airshed Model currently requires data on hydrocarbon emissions divided into five categories:

- > Paraffins
- > Ethylene
- > Olefins excluding ethylene
- > Aromatics
- > Aldehydes and ketones.

Other photochemical models might require somewhat different divisions. When such data are not available for all source types, either a special study must be undertaken or estimates must be made based on surveys in other areas, such as that reported by Trijonis and Arledge (1975). However, it would be fruitful to review the sensitivity of the model predictions to the level of detail in these inventories before embarking on a potentially costly data collection program.

c. Air Quality Data

Air quality data are required in a model evaluation study for two purposes: (1) to define background or upwind pollutant concentrations* to which emissions from the sources of interest are added, and (2) to provide the basis for comparison with model predictions. Unfortunately, data obtained for the latter purpose are of limited use for the former. The air quality measurements most valuable for comparison with model predictions are generally those taken in the areas of highest pollutant concentrations in the modeling region, but these measurements are of little value in determining background concentrations.

---

* In some models, it is also necessary to specify initial pollutant concentration conditions throughout the modeling region.

Thus, the air quality data needs of the evaluation program are dependent not only on the model being evaluated, which specifies the background or boundary and initial conditions data needed, but also the performance measure to be used, which dictates the type and amount of air quality data needed for comparison with predictions. Information on required data for various performance measures is contained in the companion report by Hayes (1979).

Data to aid in the specification of background, initial, or boundary conditions are needed for all types of models. For a Gaussian single source model used for an inert pollutant, only information on the background concentration is required. If the facility being modeled currently exists, background measurements should be made in such a manner that they are not influenced by emissions from that facility. Such measurements should be made under all conditions to which the model will be applied. In the case of a multiple-source model, the same principles apply, but additional monitoring may be necessary to characterize the upwind background concentrations adequately. The number of monitors required to obtain this information is related to the spatial variability of the upwind pollutant concentration field.

Grid models, which need initial conditions for every cell, and boundary conditions, throughout the period being modeled, require much more air quality data than do multiple-source Gaussian models. For the specification of initial conditions, an array of surface observations is required. Vertical pollutant soundings are desirable to characterize concentrations aloft. When these data cannot be collected, then concentrations aloft must be estimated based on the results of other pertinent studies or inferred from the ground-level observations. Hayes, Reynolds, and Roth (1977) suggested that up to ten stations making continuous measurements was a desirable level of monitoring for ground-based observations of air quality, while at least three and up to eight vertical sounding sites making up to eight measurements per day would be needed to characterize concentrations aloft adequately. Of course, these inputs can be specified using fewer observations, with possible attendant increases in their uncertainties. These numbers, suggested for an urban area of 1 million people and a 2500 km^2 area, were based on experience with the SAI Airshed Model.

d. Data Quality

Under this heading we consider all issues that relate to the suitability of the data employed in the model evaluation effort. We divide these considerations into three broad categories:

> Resolution of the data
> Area and time period covered by the data
> Precision and accuracy of the data.

1) Resolution of the Data

Ideally, data should be spatially and temporally resolved to the degree required by the model. For techniques that utilize relatively simple algebraic expressions to calculate concentrations, such as Gaussian models, point measurements like those obtained using a monitoring network are appropriate. That is, the Gaussian formula can be used to estimate the concentration at the exact point where the monitoring instrument is located. Care should be taken, however, that the pollutant concentrations measured for comparison with predictions are due solely to the emissions source being modeled. Because Gaussian models are steady-state models, they do not predict any temporal concentration fluctuation; therefore, the time-averaging of air quality data should be commensurate with that of the meteorological inputs.

For models that compute a concentration that is representative of the spatial average over a cell, such as grid models, there is always a question about the degree to which pollutant concentrations measured at a point in the cell represents the average concentration in the entire cell. If the measurement is not representative, it cannot be readily compared with the computed value, and thus the model's performance cannot be properly evaluated. Specific criteria governing monitor placement to ensure representativeness are not available, but as a general principle it is undesirable to locate the monitor in an area where pollutant concentration gradient may be large. For example, it would be advisable to avoid placing a monitor near a significant source of any pollutant of interest. The averaging time relevant to pollutants simulated by this type of model is usually one hour, and since the time steps used in model calculations are of the order of a few minutes, computed concentrations can be time-averaged and compared with the observations.

2) Area and Time Period Covered by the Data

One aspect of data quality that must be given attention in the application of models with large data requirements is the adequacy of coverage in a spatial and temporal sense. In most applications it seems inevitable that some compromises in the amount of data collected will need to be made, because of lack of resources, or time, or both. As was pointed out above, it is not possible with current methods to design objectively a data-gathering network. Thus, it is usually necessary to rely in part on experience and judgment to specify the number and locations of monitors. Generally, data collection for model evaluation must be more extensive than for model exercise (Hayes, Reynolds, and Roth, 1977).

In view of the probability that the input data base will prove to be less comprehensive than is ideal, the sensitivity of the model to degradation of data should be known. This sensitivity can be studied as part of model development activities, or it can be determined by making model computations with various levels of detail in an existing comprehensive data base, such as that collected in the RAPS program for St. Louis. For instance, Table IV-7 shows three possible levels of detail for the data base for input to the SAI Airshed Model (Tesche, 1978). Evaluation of model results using these different data bases should enable determination of the kinds of uncertainties that could be introduced in model predictions using less than ideal data bases.

TABLE IV-7. LEVELS OF DETAIL IN DATA USED AS INPUT TO PHOTOCHEMICAL AIR QUALITY SIMULATION MODELS

Input	Maximum Practical Level	Commonly Used Level	Minimum Acceptable Level*
Atmospheric stability	Continuous monitoring of mixing depths with acoustic sounder at one or more locations Several (5-8) vertical temperature soundings throughout the day at various locations within the modeling region Numerous surface temperature measurements recorded hourly at various locations throughout the modeling region One or more instrumented towers providing continuous measurements of the mixed layer thermal structure	A few (3-5) temperature soundings at different times of the day at one or two locations Several surface temperature measurements recorded at various locations throughout the modeling region	Twice daily temperature soundings at an airport within or nearby the region being modeled A few (1-3) surface temperature measurements with which to estimate temporal variation Limited spatial resolution or none at all
Wind fields	Numerous ground-based monitoring stations reporting hourly average values Frequent upper air soundings at several locations throughout the modeling region Continuous upper level measurements on one or a few elevated towers Wind, inversion, temperature, and terrain data used as input to the 3-D numerical model yielding the mass conserving 3-D wind field	Interpolations from ground-based monitoring network and limited (3-5) number of upper level soundings at one or two locations Resultant wind field rendered mass consistent by divergence-free algorithm	Interpolations from limited (3-5 stations) routine surface wind data; theoretically derived vertical profile assumed
Solar radiation	Several (3-5) UV pyranometers located in the region, continuously recording UV radiation levels Vertical attenuation of radiation at a few locations several times daily determined by aircraft observations Spatial (3-D) insolation fields determined by interpolation of measurements	A single, ground-based net radiometer; insolation assumed constant over the region Vertical attenuation estimated empirically as a function of aerosol mass	No radiation measurements available; estimated theoretical values based on the solar zenith angle Attenuation not accounted for
Boundary and initial conditions	Hourly species concentrations extrapolated and interpolated throughout the region using data from the extensive ground-based monitoring network; airborne data also available; hydrocarbon mix obtained from gas chromatographic analyses at several times during the day Sulfate concentrations available on an hourly basis at several locations	Hourly concentrations extrapolated and interpolated using data from several ground-based stations; hydrocarbon mix obtained from gas chromatographic analysis at one or two stations one or a few times during the day Sulfate concentrations based on a daily average and diurnal ozone curve	Hourly concentrations extrapolated and interpolated from a minimal routine monitoring network; either hydrocarbon mix assumed or average value obtained from a compilation of available data taken in a similar area No data on concentration variations aloft Sulfate measurements inferred from values obtained in similar areas

MICROFILMED FROM
BEST AVAILABLE COPY

TABLE IV-7 (Concluded)

Input	Maximum Practical Level	Commonly Used Level	Minimum Acceptable Level*
Stationary source emissions	Separate gridded inventories for point and area stationary sources; characterization of organic composition, and NO/NO_2 and SO_2/SO_4 emissions rates for major sources; diurnal and seasonal variations in nominal emissions rates for each major source type	Lumped, gridded inventory for stationary sources; NO species fractionation; seasonal and diurnal variation in regional emissions for each pollutant	Lumped stationary source emissions inventory for the region as a whole; limited information on the percentage of each source type; no temporal variation
Hydrocarbon species distribution	Mix obtained from gas chromatographic analysis of samples collected throughout the region, particularly near large sources Cold start factors applied grid by grid when calculating mobile source emissions	Mix obtained from standard emissions factors (AP-42) together with a detailed source inventory, supplemented with one or two gas chromatographic analyses	Mix assumed or obtained from available data compilation, either for the city of interest or some similar area
Mobile source emissions factors	AP-42 (latest supplement) emissions factors used in conjunction with local vehicle age distribution; corridor-by-corridor VMT, including peak and off-peak speed distributions, vehicle mix, and traffic data for intrazonal trips	AP-42 emissions factors, assumed vehicle mix, and intrazonal VMT; estimated peak and off-peak speeds, fewer traffic counts available for verification, VMT available for fewer major arterials	Gridded VMT, emissions factors estimated from 49 state mix, and average (FDC) driving profile; assumed regional speed distribution
Vehicular cold start distribution	Spatial and temporal distributions of cold starts inferred from actual traffic and demographic data	Cold starts temporally resolved using traffic distribution; no spatial resolution or spatial resolution only from estimates of driving patterns	Cold starts as a fixed percentage of all driving--traffic data are not detailed enough for spatial resolution of cold starts; cold starts estimated from demographic data
Model verification data	Hourly averaged species concentrations for NO, NO_2, O_3, SO_2, NMHC, sulfate, CO, and particulates from an extensive ground-based monitoring network	Hourly averaged concentrations of NO, NO_2, O_3, SO_2, NMHC, CO, and particulates from several ground-based stations Daily averaged sulfate measurements available from a limited (3-5) number of stations	Hourly averaged concentrations of NO_x, O_3, THC, SO_2, and CO from a minimal routine monitoring network

---

* Using data at this level of detail necessitates numerous assumptions.

Source: Tesche (1978).

In this context, we believe that it would be desirable for the EPA to consider maintaining a set of data bases suitable for evaluating air quality models. These data bases could be used for general evaluation of models and for studying model sensitivity. They could be assembled for many typical model applications, such as to urban areas (e.g., St. Louis) and to isolated point sources in flat terrain and complex terrain. The data bases could be updated as new and better information became available.

3) Precision and Accuracy of the Data

The third type of data quality issue is the precision and accuracy of the data used in model evaluation. By "precision" we mean the uncertainty of a datum about its stated value, and by "accuracy" we mean the bias in the stated value--the difference between it and the true value. We do not provide an extended discussion of these topics here since they are amply covered in the standard statistical literature. Meteorological emissions and air quality data are all subject to error. Table IV-8 shows the accuracy of meteorological measurement devices, and Table IV-9 and Exhibit IV-2 present the published precision required of federal reference methods for analyzing air pollutants. More specific information on measurement method precision can be found in many sources (e.g., Lawrence Berkeley Laboratory, 1976; Burton et al., 1976; Roth et al., 1975).

The use of an acceptable and sufficiently precise measurement technique does not guarantee the collection of high quality data. Proper data quality control procedures are also required, including the following activities:

- Procurement of adequate auxiliary equipment and supplies
- Calibration procedures
- Sampling and analysis procedures
- Data collection and reporting
- Calculation and data processing
- Preventive maintenance
- Data auditing procedures.

TABLE IV-8. VALUES AND SOURCES OF ERRORS IN METEOROLOGICAL MEASUREMENTS

Variable	Measurement Method	Scale Limitation (wavelength)	Source of Error	Approximate Accuracy
Wind direction at ground level	Vane	Micro-meso scale $\lambda = 10$ ft.	Overshoot system resonance, calibration	±5°
	Two or three orthogonal hot wires or films	Wind tunnel turbulence scale $\lambda = 10^{-4}$ ft.	Electronics calibration and response time, system reactance, ambient temperature drift	±3°
	Two or three sonic anemometers	Micro-scale $\lambda = 10^{-1}$ ft.	Electronics calibration pressure and temperature drift	±8°
Wind speed at ground level	Three-cup anemometer or propeller	Micro-meso scale $\lambda = 10$ ft.	Systems inertia, spindown time (distance constant) calibration	±0.5 mph
	Hot wire or hot film	Wind tunnel turbulence scale $\lambda = 10^{-4}$ ft.	Electronics calibration and response time, system reactance, ambient temperature drift	±1.0 mph
	Sonic anemometer	Micro-scale $\lambda = 10^{-1}$ ft.	Electronics calibration pressure and temperature drift	±1.0 mph
Temperature at ground level	Electronic (sonic thermistor, etc.)	Microscale-micro-mesoscale $\lambda = 10^{-1}$ - 10 ft.	Calibration, response lag	±3°C
Mixing or inversion height	Radiosonde	Mesoscale $\lambda = 10^4$ ft.	Tracking errors (±0.28°)	±50 ft.
Vertical temperature structure	Radiosonde	Mesoscale $\lambda = 10^4$ ft.	Sensing error	±0.5°C
Pressure	Radiosonde	Mesoscale $\lambda = 10^4$ ft.	Sensing error	±1.5 mb
Vertical wind structure	Radiosonde	Mesoscale $\lambda = 10^4$ ft.	Tracking error (±0.28°)	±4 mph ±0.7°

Sources: Data on scale limitation from Mazzarella (1972); data on source or error from MacCrady and Jex (1967). Approximate accuracies taken from those references and also MRI (1975), WMI (1974), and U.S. Army (1960). Data on radiosonde measurements from Lenhard (1973).

TABLE IV-9. UNCERTAINTIES IN MEASUREMENTS OF POLLUTANT CONCENTRATIONS BY FEDERAL REFERENCE METHODS

Pollutant	Reference Method	Precision and Accuracy
SO_2	Pararosaniline	4.6% relative standard deviation at 95% confidence level
Particulates	High volume sampler	Repeatability 3.0%, Reproducibility 3.7% (based on collaborative testing). Accuracy: Error may be as high as ±50% of the measured concentration
Carbon monoxide	Nondispersive infrared spectroscopy	Not given (analyzer must meet specifications given in Exhibit IV-2).
Photochemical oxidants	Ethylene Chemiluminescence	Not given (analyzer must meet specifications given in Exhibit IV-2).
Hydrocarbons (corrected for methane)	Flame ionigation detector	Precision: 0.5% of full scale Accuracy: 1% full scale for higher ranges; 2% full scale for lower ranges
Nitrogen dioxide	Colorimetric	Relative standard deviations: 14.4% at 140 $\mu g/m^3$ NO_2; 21.5% at 200 $\mu g/m^3$ NO_2

Source: 40 CFR §50, Appendices A-F (1975).

Performance parameter	Units [1]	Sulfur dioxide	Oxidants	Carbon monoxide	Definitions and test procedures
1. Range	Parts per million	0-0.5	0-0.5	0-50	§ 53.23(a).
2. Noise	do	0.005	0.005	0.50	§ 53.23(b).
3. Lower detectable limit	do	0.01	0.01	1.0	§ 53.23(c).
4. Interference Equivalent					§ 53.23(d).
Each interferent	do	±0.02	±0.02	±1.0	
Total interferent	do	0.06	0.06	1.5	
5. Zero drift, 12 and 24 hr	do	±0.02	±0.02	±1.0	§ 53.23 (e).
6. Span drift, 24 hours					
20 percent of upper range limit	Percent	±20.0	±20.0	±10.0	§ 53.23(e).
80 percent of upper range limit	do	±5.0	±5.0	±2.5	§ 53.23(e).
7. Lag time	Minutes	20	20	10	§ 53.23 (e).
8. Rise time	do	15	15	5	§ 53.23 (e).
9. Fall time	do	15	15	5	§ 53.23 (e).
10. Precision					§ 53.23 (e).
20 percent of upper range limit	Parts per million	0.01	0.01	0.5	
80 percent of upper range limit	do	0.015	0.01	0.5	

[1] To convert from parts per million to microgram per cubic meter at 25° C and 760 mm Hg, multiply by M/0.02447, where M is the molecular weight of the gas.

Source: 40 CFR §53.20 (1975).

EXHIBIT IV-2. PERFORMANCE SPECIFICATIONS FOR AUTOMATED MEASUREMENT METHODS

These items are discussed in the EPA's "Quality Assurance Handbook for Air Pollution Measurements" (EPA, 1976). All data collection should at a minimum fulfill the requirements of these guidances.

5. Assessment of the Need to Collect Additional Data

As a result of the previous activities, the group responsible for model evaluation now knows which pertinent existing data are available for the study and what specific data are still needed. In principle, any identified data requirement that cannot be met from previous work will need to be collected.

It should be noted that although much of the data in the emissions inventory is independent of meteorological conditions, the same is not true of air quality data. However, those elements of the emissions inventory that are dependent on weather (e.g., space heating and electrical power generation emissions depend on ambient temperature) should be obtained for conditions corresponding to those used for the modeling study. Most diurnal and seasonal variations in emissions, except diurnal traffic patterns, are neglected in practice. Air quality observations, of course, depend on meteorological conditions in a very direct way, and thus the air quality data and meteorological data should be obtained for the same day.

This stage of the model evaluation study is thus concerned with delineating additional information required for the emissions inventory and ascertaining whether sufficient corresponding meteorological and air quality data exist to evaluate the model properly. It will probably be found that, for most areas, the emissions, meteorological, and air quality data requirements will be partially fulfilled by the existing data base. The prime concern will be to delineate any supplementary data-gathering efforts needed to assemble a minimally acceptable data base for model evaluation. Of course, as pointed out above, only certain types of data can be added in a supplementary effort.

6. Specification of Performance Standards and Measures

The rationale behind and the process of specifying performance standards and measures are fully described in the companion volume (Hayes, 1979). Table IV-10 presents the model performance measures and standards developed in that work, which is one of the most systematic considerations of measuring model performance carried out to date. While four generic types of performance measures were identified--peak measures, station measures, area measures, and exposure/dosage measures--the difficulty of measurement and the unreliability of all but station measures leaves station measures as the only practical candidate for measuring model performance. The importance of station measures in evaluating model performance emphasizes the necessity of choosing locations of monitoring stations carefully.

C. IDENTIFICATION OF THE SCOPE AND REQUIREMENTS OF MODEL EVALUATION

At this point in the model evaluation study, all of the planning work has been completed. Before the execution of the study, a detailed description of the required tasks should be assembled. Such a description will ensure that nothing has been left out, that the various efforts are coordinated, and that necessary resources are allocated. Items that should be included for this definition phase are:

- > Delineation of required model characteristics.
- > Description of any necessary model modifications.
- > Listing of available resources.
  - Money.
  - Manpower.
  - Equipment.
  - Data.
- > Description of requirements for further data collection.
- > Description of the analyses to be conducted with the model results.
- > Description of the performance standards and measures to be used.

TABLE IV-10. MODEL PERFORMANCE MEASURES AND STANDARDS

Performance Attribute	Performance Measure	Performance Standard
Accuracy of the peak prediction	Ratio of the predicted station peak to the measured station (could be at different stations)	Limitation on uncertainty in aggregate health impact and pollution abatement costs[*]
	Difference in timing of occurrence of station peak[*]	Model must reproduce reasonably well the phasing of the peak--say, ±1 hour
Absence of systematic bias[§]	Average value and standard deviation of the mean deviation about the perfect correlation line, normalized by the average of the predicted and observed concentrations, calculated for all stations during those hours when either the predicted or the observed values exceed the NAAQS	No or very little systematic bias at concentrations (predictions or observations) at or above the NAAQS; the bias should not be worse than the maximum bias resulting from EPA-allowable calibration error (-8 percent is a representative value for ozone); also, the standard deviation should be less than or equal to that of the difference distribution between an EPA-acceptable monitor[**] and an EPA reference monitor (3 pphm is representative for ozone at the 95 percent confidence level)
Lack of gross error[§]	Average value and standard deviation of the absolute mean deviation about the perfect correlation line, normalized by the average of the predicted and observed concentrations, calculated for all stations during those hours when either the predicted or the observed values exceed the NAAQS	For concentrations at or above the NAAQS, the error (as measured by the overall values of the average and standard deviation of the absolute mean normalized deviation about the perfect correlation line) should not be worse than the error resulting from the use of an EPA-acceptable monitor[**]
Temporal correlation[†]	Temporal correlation coefficients at each monitoring station for the entire modeling period and an overall coefficient averaged for all stations	At a 95 percent confidence level, the temporal profile of predicted and observed concentrations should appear to be in phase (in the absence of better information, a confidence interval may be converted into a minimum allowable correlation coefficient by using an appropriate t-statistic)
Spatial alignment	Spatial correlation coefficients calculated for each modeling hour considering all monitoring stations, as well as an overall coefficient average for the entire day	At a 95 percent confidence level, the spatial distribution of predicted and observed concentrations should appear to be correlated

---

[*] There is deliberate redundancy in the performance measures. For example, in testing for systematic bias, the mean and standard deviation of the mean (signed) deviation are calculated. The latter quantity is a measure of "scatter" about the perfect correlation line. This is also an indicator of gross error and should be used in conjunction with the estimates of the mean and standard deviation of the absolute mean deviation about the perfect correlation line.

[§] These measures are appropriate when the chosen model is sued to consider questions involving photochemically reactive pollutants subject to short-term standards.

[†] These may not be appropriate for all regulated pollutants in all applications. When they are not, standards derived based on pragmatic/historic experience should be employed.

[**] By "EPA-acceptable monitor" we mean a monitor that satisfies the requirements of 40 CFR §53.20.

Source: Hayes (1979).

MICROFILMED FROM
BEST AVAILABLE COPY

This task represents the transition between the planning and the execution of the model evaluation study. The formal description of the complete model evaluation study will permit the identification of possible flaws in the modeling approach and compromises to be made in evaluating model performance.

D. PERFORMANCE OF THE MODEL EVALUATION

1. Adapting the Model for Use in the Study

This task consists of those operations necessary to prepare the model so that the data can be input to the model and translated into a set of concentration predictions for the area of interest. The operations involved are:

> Installation of the model on a computer
> Checking of model operation using test cases
> Selection of optional model features
> Adaptation of algorithms to specific area and application.

We orient our discussion to air quality models that require a computer because all models except some of the simplest exist in the form of computer programs. The first operation, that of installing the model program on the computer, can be easy if a version of the model that is compatible with the particular machine exists. If it does not, advice from computer experts should be sought to estimate the time and effort required to convert the program. These estimates should be compared with the cost of acquiring time on a computer of a more suitable type, so that a decision can be made as to where to run the program. Model developers should attempt to make their computer codes as portable as possible to ensure that a reasonable choice of computer hardware options is available to the user.

After the program is installed and running, its operation with actual data should be verified if possible. If the program has been used before, input data from a previous successful run can be used for check-out. It is important at this stage to examine all parts of the program operation, so that one can be confident that the codes are properly installed in the host computer.

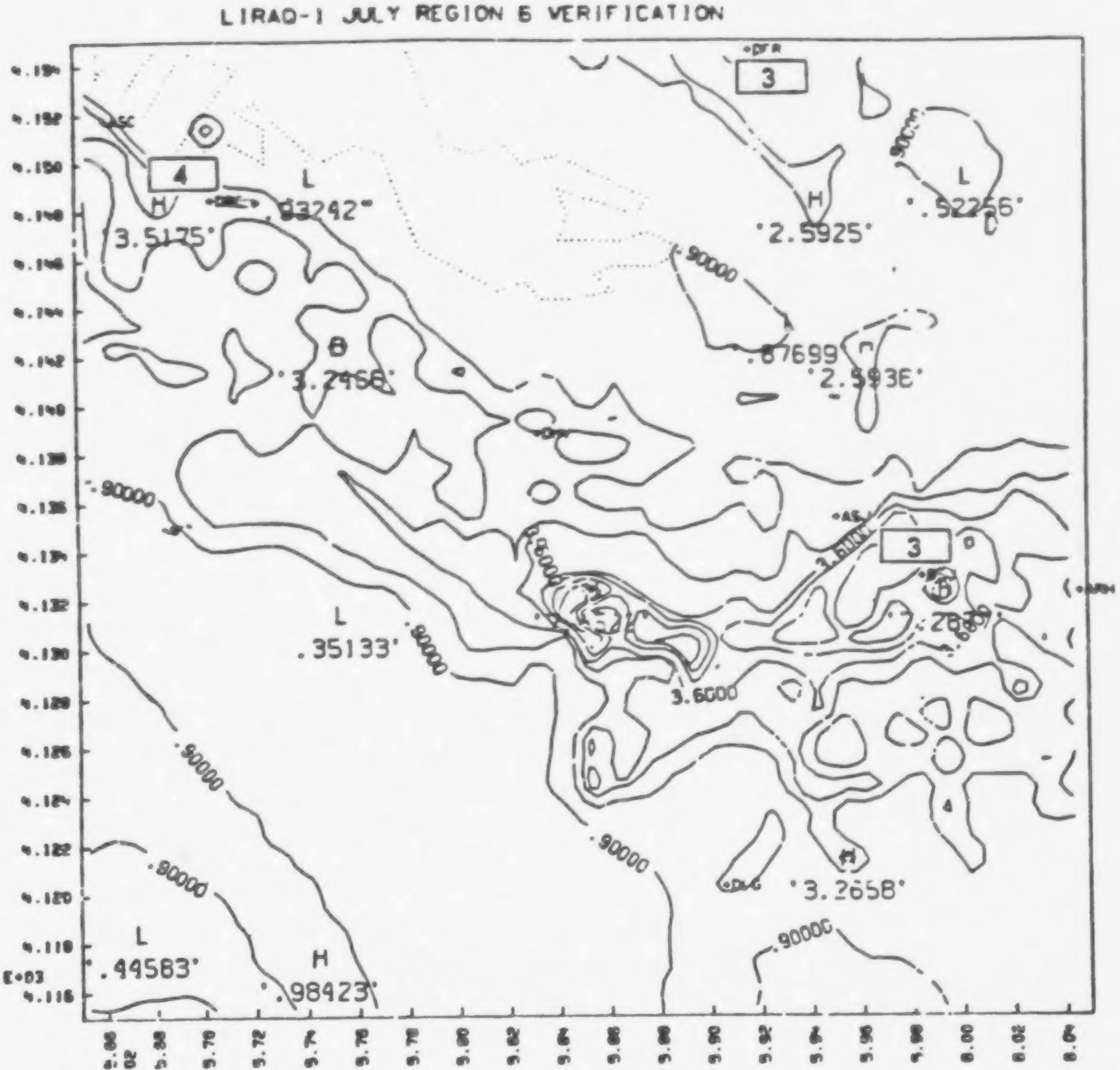
Another part of adapting a model for use in the contemplated study is to modify the model to cover the area of interest adequately and to treat the physical and chemical phenomena of importance in the study area. For instance, the coordinates of the boundaries of the modeling region must be included in the input for a model that predicts a concentration field over a given area. For the Climatological Dispersion Model (CDM), a long-term Gaussian model, this amounts to specifying the coordinates of a set of receptors at whose position the pollutant concentrations will be computed. For a grid model, the boundaries of the modeling region must be chosen, as well as the grid size. Smaller grid sizes give more resolution of output but take more computer time. (Of course, the resolution of the output is also dependent on suitable resolution of the input data.)

To illustrate the influence of grid size on model output, we present in Figure IV-3 surface concentration isopleths calculated by the LIRAQ Model (MacCracken et al., 1977) for CO using a 1-kilometer grid and a 5-kilometer grid, respectively. There is much more detail in the 1-kilometer grid results than in the 5-kilometer results. Comparison of the two figures shows the same broad trends, though calculation using the coarser grid produces lower predictions of peak concentrations.* The size of the grid used should be compatible with the resolution of the available input data as well as the desired resolution of the output. In general, a grid model is capable of resolving only features in the concentration field that have spatial scales of at least two grid cell dimensions.

Finally, it is necessary to adapt model algorithms for the application. For instance, any model that deals with point sources has embedded in it a

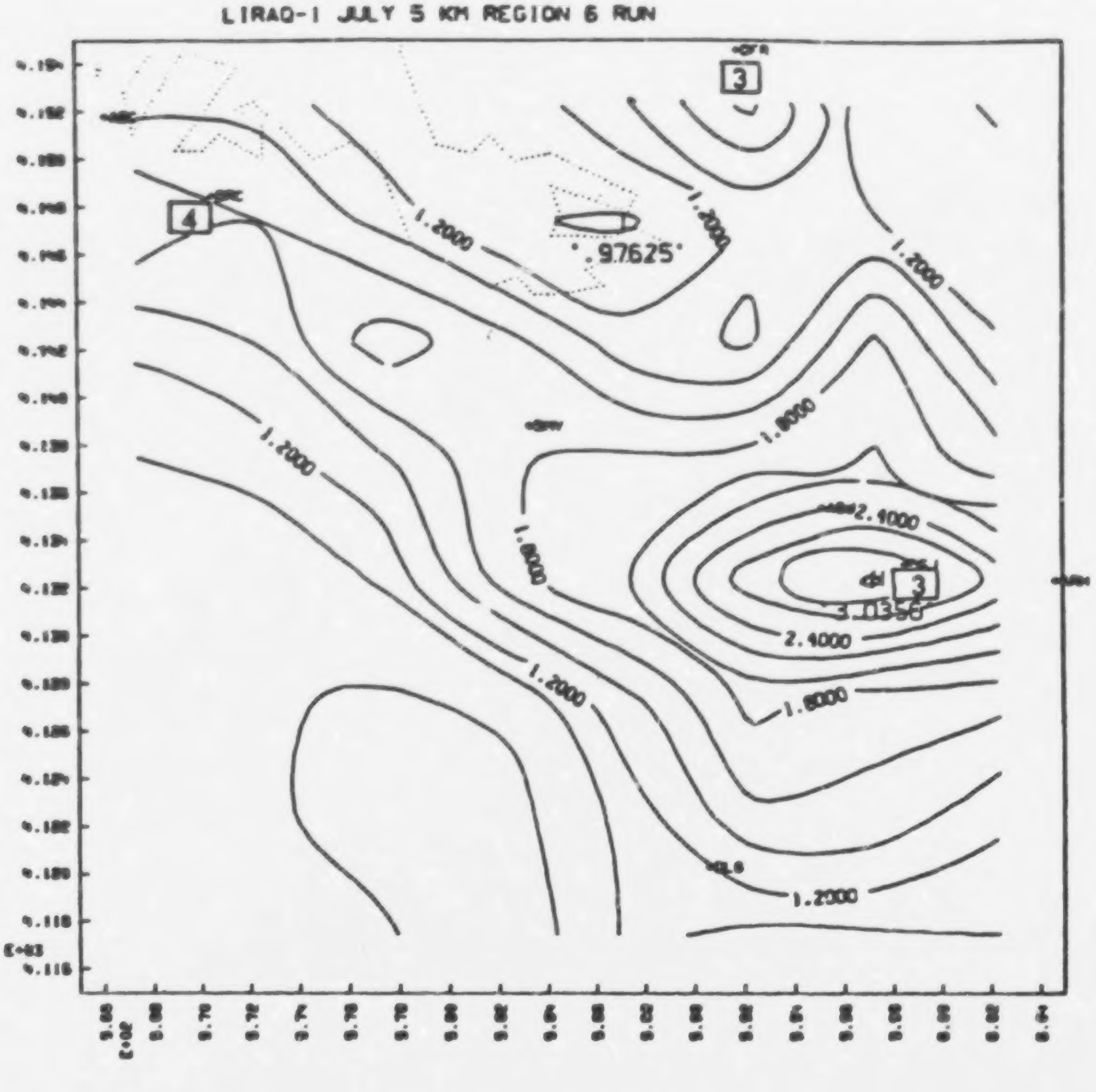
---

* We note that the lower values of the peak predicted concentrations are expected because the model predictions represent an average over a much larger area when the 5 kilometer grid is employed. This has the effect of averaging out the peak values.


LIRAQ-1 JULY REGION 6 VERIFICATION
4
3
3
H
'3.5175'
L
.93242"
H
'2.5925'
L
'.52256'
.90000
'3.2458'
.87699
'2.5938'
'.90000
3.6000
L
.35133'
3.6000
.90000
.90000
.90000
'3.2658'
.90000
L
'.44583'
H
'.98423'
E+03

(a) 1-Kilometer Grid

FIGURE IV-3. SURFACE CO CONCENTRATIONS CALCULATED FOR THE SAN FRANCISCO BAY AREA AT 1400 ON 26 JULY 1973 USING THE LIRAQ AIR QUALITY MODEL AT DIFFERENT GRID SIZES



(b) 5-Kilometer Grid

Source: MacCracken et al. (1977).

FIGURE IV-3 (Concluded)

provision for incorporating a plume rise algorithm. Many plume rise algorithms have been proposed in the technical literature. Because there is some latitude in the selection of a plume rise algorithm for a given application, the choice should involve comparison of computed values by different algorithms with observations. Other algorithms that may need adaptation for a particular situation are:

- The technique used to construct wind inputs given the available meteorological observations.
- The method for estimating mixing depths.
- The turbulent dispersion or diffusivity algorithm.
- The kinetic mechanism employed to treat chemical reactions.
- The algorithms for generating emissions inputs.

In general, the user should examine all algorithms in the model in light of the current understanding of physical and chemical processes known to be important in the study area.

2. Gathering, Assembling, and Formatting the Required Data

We discuss gathering, assembling, and formatting data together, since they are obviously highly interrelated.

As the data base is assembled, it is codified and formatted for use by the computer program. The first part of this task consists of collecting existing data from emissions inventories and meteorological and air quality data bases, as discussed in Section IV.B.3. Then additional data, as discussed in Section IV.B.5, may need to be gathered. For the emissions inventory, the requisite traffic surveys should be undertaken, and necessary data pertaining to industrial activities and land use should be collected. For the meteorological and air quality data collection, supplemental monitoring should be carried out for the necessary periods of time.

Data assembly and formatting include the following activities:

> Evaluate and verify the data
> Load the data on the computer
> Prepare the data bases for model use.

Data evaluation and verification involve examination of the collected data to ensure that adequate quality control procedures were applied in their collection. The data validation procedures discussed in the EPA's Quality Assurance Handbook (EPA, 1976) should be followed. For simple models such as the Gaussian, loading the data and preparing them for the model can entail just punching a few computer cards. In the case of a large, complex model, the data are entered into the computer and then used as input to a set of data preparation programs, which prepare the necessary data files for use with the simulation program. These data files can be checked for errors and consistency before the main program is exercised. At the conclusion of this phase of the evaluation, a complete data base has been assembled for exercising the model under the appropriate conditions.

3. Exercising the Model

In this task, the computer runs necessary to evaluate the model are carried out. If all of the preceding tasks have been executed properly, no further difficulties should be encountered. All results of the model runs should be saved on computer files for later access.

At this point, selected performance measures will be calculated (as discussed in Section IV.B.6). Although these measures will be the means whereby the model will be judged to be adequate or not, all other results produced by the model runs must be saved since they will be needed for further analysis should model performance be unsatisfactory.

4. Analyzing the Results of the Evaluation

The first step in analyzing the results of the model evaluation is to compare each computed performance measure with the appropriate performance standard. If the measure meets the standard, the model is considered verified, and one can conclude that model performance is satisfactory for the application. Thus, the user can proceed with some confidence that the verified model will produce reliable results for the intended application.

If a performance measure fails to meet a performance standard in some respect, the model is not considered verified and further analysis is indicated. (Even if the measure meets the performance standard, further analysis as outlined below will give much useful information about model behavior and is worthwhile to carry out if possible.)

The analysis of the model evaluation results should center on the differences between computed and observed pollutant concentrations, i.e., the residuals. These residuals can arise from three sources:

- Errors in the input data (emissions data, meteorological data, initial and boundary conditions, and background concentrations).
- Errors in the formulation of the model (approximations made in modeling pollutant dispersion, transport, or chemical transformations).
- Errors in the air quality measurements used for comparison.

The magnitudes of the first and third sources of error can be estimated from the characteristics of the methods by which they are measured, using standard statistical techniques. In addition to measurement errors that result from statistical fluctuations in measurement devices, nonzero residuals can result, requiring the extrapolations and interpolations necessary to generate a complete model input data set from insufficient data. The errors introduced by these extrapolations and interpolations are more difficult to quantify than instrumental errors. They can be investigated by carrying out an evaluation of the interpolation algorithm similar in structure to that described earlier for the full air quality model.

Discrepancies introduced by shortcomings in the model's formulation are difficult to evaluate because we have no "true" model of the relevant atmospheric processes with which to compare it. Also, in light of the uncertainties of the input and comparison data, error due to model formulation cannot be isolated from the total modeling process.

The air quality data used for comparison with model outputs introduce an uncertainty that can be quantified. Because this uncertainty stems from the measurement process, it can be determined by replication of measurements and standard statistical anlaysis.

Careful analysis of the residuals can yield much useful information about the model, even if quantitative statements about sources of error cannot be made. Several different ways of analyzing residuals can be informative [see, for example, Koch and Thayer (1971)]. Plots of residuals against time of day can reveal systematic biases, which might result from one of a number of assignable causes, such as an inadequate kinetic mechanism in a photochemical model. Dependence of the magnitude of residuals on concentration might indicate that a monitoring location is poorly placed to detect a large area-wide concentration level, or that the wind field generated by the model mislocates the plume from a particular source. Differences between residual dependencies for primary and secondary pollutants can be used to infer deficiences in the kinetic mechanism or dispersion processes. The examples given here are only a sampling of possible situations; this type of analysis should be guided by the particular situation and a knowledge of the model's characteristics.

Other statistically based analysis methods that can be used to study the evaluation results are:

> Plots of residuals against exogenous variables.
> Scatter plots of observed and computed concentrations.
> Correlation between observed and computed concentrations.
> Nonparametric tests of location to indicate possible bias of computed concentrations relative to observations.

5. Assessing the Need To Perform Further Evaluation

Upon completion of the analysis of the results of the evaluation, some indications of possible inadequacies in the model should be apparent. This diagnosis can result in various actions:

- Reformulate model. Instead of switching to a completely different model, consideration can be given to reformulating parts of the chosen model. For instance, the plume rise algorithm can be changed, or an interpolation algorithm for input data can be altered.
- Carry out additional verification. If a performance measure upon evaluation does not seem appropriate, alternative choices could be identified and evaluated. In addition, consideration could be given to examining more cases.
- Collect more data. If deficiencies in the input data are identified, supplementary data collection should be instituded and the model should be reevaluated. Further data collection can also augment air quality data used for comparison purposes and to evaluate the performance measures.
- Reexamine the choice of model. If it is shown that the model being considered is inadequate for the purpose and no other recourse is possible, consideration could be given to using a different model.

It is likely that the data already collected would support a subsequent evaluation study of a new or reformulated model. Thus, the time and effort that went into the original evaluation study would by no means be wasted even if the initially chosen model is found to perform inadequately.

6. Evaluating the Adequacy of the Model

Basically, there are two possible outcomes of the evaluation: The model is satisfactorily verified or it is not. In the former case, the model can be used for the application with some confidence that it will produce satisfactory results. In addition, if the analysis of the evaluation study results was done thoroughly, much information was gained as to the operation of the model and the situations for which it might not give acceptable results.

In the case where the model was not verified, the preferable action is to reconsider the use of the model in the form tested. Action to rectify deficiencies should be taken before the model is applied in further studies. However, circumstances may dictate that, even though unverified, the model must be used anyway. For example, the analysis may be subject to deadlines imposed by governmental regulations. Perhaps the model represents the current state of the art and no improvements are possible without a major research effort. Whatever the reason, if the model is used, the previous analyses may pinpoint some of the deficiencies in the model that caused it to fail to meet the required performance standards. These deficiencies should be fully detailed in the account of its use. To provide some assistance and guidance to the model user under such circumstances, we suggest that a central model evaluation group be formed. The model user could seek help from this group should the model evaluation be unsuccessful.

E. EVALUATION FOR SCREENING APPLICATIONS

Air quality model usage can be segmented into screening and refined applications. As an example of the former, a relatively simple model might be used to identify the potential existence of an air quality problem. If a more detailed or refined analysis is contemplated, then a more sophisticated model would be employed to achieve the requisite accuracy in the computed results.

Although we have discussed model evaluation in the context of refined usage, this does not imply that all screening applications do not merit or require some appropriate type of evaluation. The performance of some possible screening models, such as rollback and the Empirical Kinetic Modeling Approach (EKMA), cannot be readily evaluated owing to [illegible]treme paucity (if not unavailability) of the requisite data.* For other screening models, the evaluation process should be limited to establishing that model performance is adequate for screening purposes. We suggest that, to the extent possible, all screening models be subjected to a comprehensive evaluation prior to their adoption for such usage. This evaluation could generally follow the guidelines described in this chapter. In addition to establishing the performance characteristics of these models, it will be necessary to identify appropriate input data requirements to ensure adequate performance.

Air quality models are made up of many component parts. Occasionally there could be a need to evaluate particular parts, or "modules," of a model. The recommendations in this chapter have been formulated with evaluation of the complete model in mind. However, there is no reason why the methods should not be used to evaluate particular modules, if the necessary data bases can be assembled. The principles of such a module-by-module evaluation are similar to those laid out here.

F. PERSPECTIVE

As we have developed in this chapter, there are many issues that need to be properly taken account of in carrying out an air quality model evaluation. To aid the reader in placing the material in this chapter in perspective, we close with a brief retrospective on the context of model evaluation studies.

It seems likely that in the future air quality model usage will be widespread. For example, the need to demonstrate compliance with schedules for improving air quality will be one problem that will require the use of modeling. Extensive model use will create a need for adequately verified models of all types. At present, there are no formal procedures specified for model evaluation, though individual groups have carried out numerous evaluation studies. These past studies have lacked common bases and goals; thus, they do not yield much information of a general nature.

---

* Plans for a study to assess the feasibility of employing relatively sophisticated air quality models to determine the performance characteristics of simpler procedures such as EKMA are described by Tesche (1978).

In the discussion in this chapter, we have attempted to lay out in some detail a conceptual framework for model performance evaluation and to identify the component parts of such a study. The current needs for model evaluation are: (1) to specify in detail all of the steps of an evaluation study, (2) to develop formal procedures and techniques where they do not now exist, (3) to prepare a guidelines document, and (4) to implement those procedures. In the final chapter we discuss some needs in the areas of institutional requirements, technical development, and documentation.

V RECOMMENDATIONS

In this report we have outlined the relevant issues in air quality model evaluation, setting out a comprehensive structure for carrying out such an evaluation. Inasmuch as this work represents a first effort to systematically address all of the relevant issues, it has provided an opportunity to discover that there are significant gaps in the background information required for successful model evaluation. In this Chapter we list items that we have identified as necessary adjuncts to establishing model evaluation as a routine element of air quality model use.

The recommendations are divided into three sections: institutional, technical development, and documentation. In the first section, we describe a set of functions for a model evaluation group that might be established by an appropriate private or governmental body. The second section contains suggestions for technical development work that might be undertaken to clarify those aspects of model evaluation in which, currently, judgment must be substituted for applicable study results. In the final section we list guideline documents that we believe should be available to the future model user who wishes to first evaluate properly the model he intends to use.

A. INSTITUTIONAL NEEDS

We believe that it would be highly desirable to set up a group of experts who would be responsible for providing assistance in air quality model evaluation within the model development and user communities. This group, which could consist of experts in air quality and many different fields, could offer assistance in carrying out model evaluations. In addition, they could call on other outside experts when necessary. The functions of such a group might include:

- Establishing guidelines and practices for all aspects of model performance evaluation.
- Supplying expertise in setting up air quality model evaluation studies and assisting in such studies.
- Coordinating (and possibly supplying) funds and equipment necessary for model evaluation work.
- Maintaining a central information exchange on the status of air quality model evaluation results.
- Developing, evaluating, and maintaining adequate and diversified data bases for future model evaluations.
- Making judgments on the need for evaluative studies in specific cases.
- Assigning responsibilities for model evaluations.
- Compiling, maintaining, and updating a set of guidelines documents covering all aspects of model evaluation.
- Developing guidelines for model selection based on evaluation experience.
- Determining the disposition of cases in which performance evaluation is unsatisfactory.

To carry out the above tasks, this model evaluation group will need to have access to air quality experts with specialized knowledge in the following fields (c.f. Chapter IV, Section B.2.b.):

- Meteorology
- Analytical chemistry
- Computer programming
- Statistical analysis
- Air quality analysis.

The group should be staffed at a level commensurate with the amount of ongoing modeling and model development activity. Some extra effort may be desirable to get initial studies underway; however, once some experience is gained, it may be possible to decrease the commitment of resources.

B. AREAS FOR TECHNICAL DEVELOPMENT

Several areas pertinent to model evaluation studies currently lack a sound and complete technical basis. Consequently, evaluation study design drawing on these areas necessarily relies heavily on the judgment of experienced scientists rather than on pertinent results. We list here those areas that we have identified, in developing our model evaluation procedures, as meriting further study.

- Determination of the circumstances under which model evaluation is not mandatory. As pointed out in Chapter IV, a model that is to be applied in a situation for which a successful evaluation has previously been carried out might not require further evaluation. The specific circumstances under which a previous evaluation would be transferable should be investigated.
- Determination of the amount and quality of data needed to carry out a model evaluation. In general, the amount of data needed to evaluate a model is greater than that necessary for a routine study. The relationship between model performance standards and measures on one hand, and adequacy of data to calculate them on the other, should be investigated.
- Definition of an appropriate period of record for stochastically varying quantities such as meteorology. This would be applicable to long-term average models.
- Development of uniform programming standards for the computer programs, to ensure orderly transfer of air quality models between different user groups.

CONTENTS

DISCLAIMER . . . . ii 1/A5

ACKNOWLEDGEMENT. . . . . iii 1/A6

LIST OF ILLUSTRATIONS. . . . . vii 1/A10

LIST OF TABLES . . . . ix 1/A12

LIST OF EXHIBITS . . . . xi 1/A14

I INTRODUCTION. 1 1/B1

  A. Purpose of This Study 2 1/B2
  B. Structure of This Report. 3 1/B3

II REVIEW OF PREVIOUS MODEL EVALUATION EFFORTS 5 1/B5

III KEY ISSUES IN THE EVALUATION OF AIR QUALITY 9 1/B9

IV A PROCEDURE FOR EVALUATING MODEL PERFORMANCE. 17 1/C3

  A. Problem Specification and Model Selection 20 1/C7
    1. Setting the Context of Model Usage. 20 1/C7
    2. Selection of a Model. 22 1/C9

  B. Planning the Model Evaluation Study 23 1/C10
    1. Determination of the Need for Model Evaluation. 23 1/C10
    2. Development of a Conceptual Evaluation Plan 26 1/C13
    3. Examination of Existing Data Bases. 37 1/D10
    4. Assessment of Data Needs. 43 1/E2
    5. Assessment of the Need To Collect Additional Data. 64 1/F11
    6. Specification of Performance Standards and Measures. 65 1/F12

  C. Identification of the Scope and Requirements of Model Evaluation. 65 1/F12

  D. Performance of the Model Evaluation 67 1/F14
    1. Adapting the Model for Use in the Study 67 1/F14
    2. Gathering, Assembling, and Formatting the Required Data 71 1/G4
    3. Exercising the Model. 72 1/G5

4. Analyzing the Results of the Evaluation 72 1/G5
5. Assessing the Need To Perform Further Evaluation. 75 1/G8
6. Evaluating the Adequacy of the Model. 75 1/G8

E. Evaluation for Screening Applications 76 1/G9

F. Perspective 77 1/G10

V RECOMMENDATIONS 79 1/G12

A. Institutional Needs 79 1/G12

B. Areas for Technical Development 81 1/G14

C. Documents To Be Compiled. 82 2/A3

D. Summary 83 2/A4

APPENDIX: SUMMARY OF PREVIOUS EVALUATION STUDIES. 85 2/A6

REFERENCES 141 2/F5

> Study of the number of monitors required for a model evaluation and their optimum placement. The results of this investigation would have wide applicability.
> Study of the costs of data collection and model simulations. Although collection cost studies have been carried out in the past, these estimates should be updated periodically.
> Study of the degree to which air quality and meteorological measurements are representative of their environment. This work would clarify the problem of comparability between observed and computed concentrations.
> Development of a set of data bases suitable for model evaluation work. A study designed to develop such a set of data bases would consider appropriately generalized scenarios and develop meteorological, emissions, and air quality information necessary for evaluating the performance of a wide variety of models.

C. DOCUMENTS TO BE COMPILED

There is a clear need for guidance in many of the necessary tasks to ensure proper evaluation of air quality models. We list here a series of guideline documents that could be developed to aid the proposed model evaluation group:

> Air quality model performance evaluation.
> Selection of model performance standards and measures.
> Air quality monitoring for model performance evaluation.
> Model selection and applications.
> Air quality model input data base preparation.

D. SUMMARY

In the above discussion, we have outlined many areas in which further research is desirable. The number of areas is an indication of the current state of model evaluation. Until now, model evaluation has typically been carried out on a more or less ad hoc basis by model developers. There is a clear need for a formal framework for model evaluation. This report represents a first step towards the development of that framework.

BLANK PAGE

APPENDIX
SUMMARY OF PREVIOUS EVALUATION STUDIES

APPENDIX

SUMMARY OF PREVIOUS EVALUATION STUDIES

In Chapter II we discussed the results of some previous model evaluation studies. This appendix gives detailed results from those evaluations, together with comments on the methods employed. We have not attempted a comprehensive survey of evaluation stuides, but rather have chosen some representative studies to illustrate the methods used. Table A-1 lists the models that are covered by the studies included.

Our review of air quality model evaluation studies has revealed that emphasis so far has been on matching pollutant concentrations predicted by a model with point measurements at monitoring stations. Measures used for comparison include:

> Correlation coefficients.
> Differences between observed and predicted values--either mean or root-mean-square differences, and either absolute or relative differences.
> The ratio of observed to predicted concentrations.
> Frequency distributions of pollutant concentrations.
> Regression statistics.
> Qualitative comparisons.

Evaluation studies have concentrated on statistical measures of agreement with station observations. Although most model developers have considered evaluation to be a part of the development process, they have in general used existing data instead of collecting new data designed specifically for evaluation purposes. No studies to date have attempted to match

TABLE A-1. MODELS CONSIDERED IN MODEL EVALUATION STUDIES DESCRIBED IN THIS APPENDIX

Model Type	Model Name	Page
Gaussian	CRSTER	100
	PTMTP	101
	CDM	94, 107
	AQDM	94
	SCIM	94
	TRAPS	110
	CALINE-2	110
	HIWAY	110
	AIRPOL-4	110
	APRAC-1A	110
Box	Gifford-Hanna	
Plume	RPM	124, 125
Grid	SAI	116, 118
	LIRAQ	121
	Shir and Shieh	126
Trajectory	DIFKIN	116
	REM	116

model results not expressed as concentrations, such as areas in violation of air quality standards or exposure/dosage information, because of the lack of appropriate observational data.

Results from the studies reviewed show performance of air quality models varying over a wide range. However, from available evaluative results, it is difficult to draw general conclusions about relative model performance that could serve as a basis for selecting a model for a particular situation. Even definite conclusions about the validity of each model considered cannot be deduced. Thus, a set of standardized guidelines for carrying out a model evaluation effort is needed. Such a set of procedures will permit reasonably definite statements to be made about model validity. Evaluation studies should produce evidence of model performance in both an absolute and relative sense. Moreover, such studies should cover a variety of conditions to test model features thoroughly.

In the review that follows, we sometimes criticize the way in which a model evaluation was carried out. This criticism, however, does not reflect on the investigators, for there were no general guidelines for model evaluation when these studies were carried out; thus, each worker of necessity designed his study to suit his own particular needs.

1. GAUSSIAN MODELS

Gaussian models are "generally considered to be the state-of-the-art techniques for estimating the impact of non-reactive pollutants" (EPA, 1977). Although a great variety of Gaussian models have been developed, they all use the same basic formulation, which assumes that steady-state pollutant concentrations downwind of a source are described by the expression:

$$\chi = \frac{Q}{\pi u \sigma_y(x)\sigma_z(x)} \exp\left\{-\frac{1}{2}\left[\left(\frac{h}{\sigma_z(x)}\right)^2 + \left(\frac{y}{\sigma_y(x)}\right)^2\right]\right\} \quad , \quad , \qquad \text{(A-1)}$$

where

χ = concentration at the receptor,
Q = pollutant emissions rate at the source,
u = wind speed,
h = effective height of the source,
y = crosswind distance between source and receptor,
x = downwind distance from source to receptor,
$\sigma_y(x)$= horizontal diffusion parameter,
$\sigma_z(x)$= vertical diffusion parameter.

With this formula as a basis, many different types of models have been developed. Some are formulated to calculate the short-term average concentration downwind of a source or group of sources (point, area, or line sources). Others, the so-called climatological models, integrate Eq. (A-1) over a long-term distribution of wind speeds, wind directions, and atmospheric stabilities, to yield long-term average concentrations over a region. Thus, Gaussian-type models are available for single sources or whole cities, and for long- or short-term average concentrations.

a. Study by Koch and Thayer (1971)

We first discuss two evaluation studies of Gaussian models reported by Koch and co-workers (Koch and Thayer, 1971; Koch and Fisher, 1973). In the first study, the objective was "to evaluate critically the predictive accuracy of the urban diffusion model based on the Gaussian plume concept." Both short-term (one- and two-hour average) and long-term (one-month and three-month average) SO_2 concentrations in St. Louis and Chicago were used for comparison with model predictions. The short-term average concentrations were calculated using a multiple source steady-state Gaussian plume model, whereas the long-term averages were evaluated using a statistically selected sample of one-hour-average concentrations.

The input data were obtained from available meteorological and air quality data. Wind speeds were obtained by averaging observations at several stations. Vertical wind profiles were obtained from a single tower in St. Louis and calculated assuming a power law wind profile for Chicago. Stability class was characterized by wind speed and radiation index, and mixing depths were interpolated from measurements 100 to 200 miles away.

The overall results for short-term average concentrations are given in Tables A-2 and A-3. For the St. Louis data, the authors commented that the mean observed and predicted concentrations at individual monitoring stations were in good agreement. However, the agreement was not as good for individual values, as can be seen by examining the standard deviations and mean absolute differences of observed minus predicted concentrations. The authors concluded that the data in Table A-1 constitute evidence of the model's ability to predict long-term rather than short-term average concentrations. Table A-2 also shows that, in general, correlation coefficients for two-hour-average concentrations were low, and the slopes of the regressions of observed on predicted values were substantially different from unity, both indicative of the generally poor agreement between observations and predictions. Similar results were observed for the Chicago data (Table A-3).

Table A-4 shows the percentages of observations for the two cities that were within various error limits. The results are generally similar--about 75 percent of the predictions are within ±(observed mean) of the observed mean. The only factor found to have a consistent effect on predictions was the wind speed. Tables A-5 and A-6 show that agreement between observed and predicted results was highly dependent on wind speed. The authors attributed this effect to inadequate estimation of diffusion parameters and/or inadequate accounting for the possible effect of wind speeds on emissions rates (e.g., fuel consumption for space heating is affected by wind speed, and the model inputs did not take this into account).

TABLE A-2. STATISTICAL SUMMARY OF OBSERVED TWO-HOUR SO_2 CONCENTRATIONS (in $\mu g/m^3$) FOR ST. LOUIS STATIONS AND CONCENTRATIONS CALCULATED USING A GAUSSIAN PLUME MODEL

Station Number	Mean			Standard Deviation			Mean Absolute Difference of Observed Minus Predicted	Regression of Observed on Predicted Values		Number of Values	Correlation Coefficient
	Observed Values	Predicted Values	Observed Mean Minus Predicted Mean	Observed Values	Predicted Values	Observed Minus Predicted Values		Slope	Intercept		
3	156	196	- 40	145	180	207	130	0.1637	123.9	1037	0.203
4	175	142	+ 33	157	195	212	116	0.2354	141.4	872	0.292
10	335	207	+128	237	165	255	201	0.3373	265.2	975	0.235
12	179	211	- 31	136	214	194	121	0.2891	118.4	980	0.455
15	137	118	+ 19	132	119	133	87	0.4964	78.6	900	0.448
17	211	181	+ 31	124	161	161	114	0.2973	157.7	1031	0.386
23	90	191	-101	106	241	238	142	0.1085	69.2	963	0.247
28	87	94	- 7	117	149	152	80	0.2849	60.1	788	0.363
33	73	61	+ 11	88	99	103	53	0.3517	51.1	922	0.396
36	80	88	- 8	78	134	122	64	0.2542	57.2	952	0.437
All	154	151	+ 3	159	179	194	112	0.3085	107.9	9420	0.347

Source: Koch and Thayer (1971).

BLANK PAGE

BLANK PAGE

TABLE A-3. STATISTICAL SUMMARY OF OBSERVED ONE-HOUR SO_2 CONCENTRATIONS (in $\mu g/m^3$) FOR CHICAGO STATIONS AND CONCENTRATIONS CALCULATED USING A GAUSSIAN PLUME MODEL

TAM* Station Number	Mean			Standard Deviation			Mean Absolute Difference of Observed Minus Predicted	Regression of Observed on Predicted Values		Number of Values	Correlation Coefficient
	of Observed Values	of Predicted Values	Observed Mean Minus Predicted Mean	of Observed Values	of Predicted Values	of Observed Minus Predicted Values		Slope	Intercept		
1	33	47	- 14	56	111	98	39	0.2349	21.6	723	0.466
2	114	99	+ 15	87	108	128	87	0.1188	102.7	602	0.148
3	312	379	- 67	152	416	397	221	0.1106	269.7	606	0.303
4	123	315	-192	89	294	274	201	0.1119	88.0	614	0.370
5	62	128	- 66	47	140	135	83	0.0936	50.2	722	0.279
6	23	58	- 35	32	98	97	45	0.0595	19.5	703	0.182
7	102	158	- 55	95	159	157	100	0.1905	72.2	711	0.319
8	43	36	+ 7	39	76	83	45	0.0366	41.8	726	0.071
All	96	145	- 49	117	232	201	99	0.2493	60.2	5407	0.494

* TAM = Telemetered air monitoring.

Source: Koch and Thayer (1971).

TABLE A-4. COMPARISON OF ERROR DISTRIBUTIONS FOR TWO-HOURLY ST. LOUIS AND HOURLY CHICAGO VALIDATION CALCULATIONS

Range of Predicted Minus Observed Concentration µg/m^3	% of Comparisons Within Error Limits	
	St. Louis (Mean Observed Concentration = 154 µg/m^3)	Chicago (Mean Observed Concentration = 96 µg/m^3)
± 5	8	8
± 10	15	17
± 20	25	30
± 50	46	53
±100	65	73
±150	76	82

Source: Koch and Thayer (1971).

When the observed and predicted SO_2 concentrations for the different stations were averaged over three months for St. Louis and one month for Chicago, improved agreement was obtained. Those results are shown in Table A-7. Averaging over all of the data at each station was intended to provide an indication of the performance of the model in predicting longer term average concentrations. However, the low values for correlation coefficients (corresponding to 46 percent of the variance explained for the St. Louis data and 76 percent for Chicago), and the slope of 0.63 for the Chicago data, indicate that agreement between observed and predicted concentrations is still not high. Koch and Thayer (1971) quoted a combined root-mean-square error (RMSE) of 68 µg/m^3, compared with an overall mean of 128 µg/m^3, indicating that, for this study, overall RMSE was about one-half of the overall mean. In addition, it was stated that the long-term mean was overpredicted at 11 stations and underpredicted at 7 stations, indicating a tendency of the model to overpredict observed concentrations more often than to underpredict. If the model actually had no tendency to overpredict or underpredict, however, the above result would be obtained with a probability of 12 percent. Consequently, that result is not a strong indicator of bias in the model.

TABLE A-5. OBSERVED, PREDICTED, AND OBSERVED MINUS PREDICTED CONCENTRATIONS BY WIND SPEED CLASS FOR ST. LOUIS DATA

Wind Speed Class (m/sec)	Two Hour Concentrations ($\mu g/m^3$)						Number of Cases
	Observed		Predicted		Obs. - Pred.		
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	St. Dev.	
1.5	171	142	659	618	-488	618	30
$1.5 < u \le 2.0$	245	200	343	250	- 98	229	110
$2.0 < u \le 2.5$	208	195	325	267	-117	281	415
$2.5 < u \le 3.0$	197	171	257	238	- 60	241	660
$3.0 < u \le 4.0$	178	162	198	210	- 21	223	1781
$4.0 < u \le 5.0$	154	140	143	151	11	169	1993
$5.0 < u \le 6.0$	136	135	113	115	24	142	1800
$6.0 < u \le 8.0$	138	180	88	93	50	164	1966
$8.0 < u \le 10.0$	104	112	71	68	33	105	600
$10.0 < u$	96	106	53	50	43	89	65

TABLE A-6. OBSERVED, PREDICTED, AND OBSERVED MINUS PREDICTED CONCENTRATIONS BY WIND SPEED CLASS FOR CHICAGO DATA

Wind Speed Class (m/sec)	Number of Cases	One Hour Concentrations ($\mu g/m^3$)					
		Observed		Predicted		Obs. - Pred.	
		Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.
$1.0 < u \le 1.5$	88	116	117	430	385	-314	356
$1.5 < u \le 2.0$	238	149	143	364	400	-215	359
$2.0 < u \le 2.5$	279	125	135	276	484	-151	450
$2.5 < u \le 3.0$	462	113	123	185	267	- 72	230
$3.0 < u \le 4.0$	1089	107	128	165	211	- 57	169
$4.0 < u \le 5.0$	1117	95	117	116	147	- 22	124
$5.0 < u \le 6.0$	929	91	112	119	159	- 28	131
$6.0 < u \le 8.0$	715	70	94	72	121	- 2	126
$8.0 < u \le 10.0$	466	55	65	53	50	1	59
$10.0 < u$	24	59	68	38	35	22	46

TABLE A-7. STATISTICS FOR LONG-TERM AVERAGE PREDICTIONS OF SO_2

City	Number of Stations	Overall Mean ($\mu g/m^3$)	RMSE ($\mu g/m^3$)	Regression of Observed on Predicted Values		Correlation Coefficient
				Slope	Intercept	
St. Louis	10	154	56	0.98	-0.56	0.675
Chicago	8	96	78	0.63	4.9	0.873

Source: Koch and Thayer (1971).

When long-term averages were calculated using a sample of the one-hour-average concentrations, the mean predicted concentrations did not change substantially, but the RMSE increased relative to the mean that was calculated using all of the data, as expected. This effect is shown in Table A-8. The authors concluded that 1 hour sampled out of 24 is as small a sample as should be used for calculating seasonal averages.

TABLE A-8. SUMMARY OF ACCURACY OF SAMPLING INTERVALS FOR ESTIMATING DISTRIBUTION OF PREDICTED CONCENTRATIONS OVER A SEASON

Sampling Interval, Hours	Root Mean Square Error (RMSE),(a) $\mu g/m^3$	Mean, $\mu g/m^3$
1	--	150
2	2.43	150
4	3.70	152
6	3.58	148
8	7.69	151
12	7.94	150
24	16.43	158

(a) $(RMSE)_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{1,i} - x_{j,i})^2}$

N = No. of Stations (10)

j = Sampling Interval

$X_{j,i}$ = Seasonal Mean Concentration for i th Station with Sampling Every Hour

$X_{1,i}$ = Seasonal Mean Concentration for i th Station with Sampling Every Hour

---

Source: Koch and Thayer (1971).

Koch and Thayer (1971) drew the following conclusions about model performance and adequacy from their study:

> Predicted long-term (monthly or seasonal) concentrations averaged over several locations are in good agreement with observed concentrations.

> Predicted long-term concentrations at individual locations show a root-mean-square error equal to about half the mean and indicate a slight tendency to overpredict observed concentrations.

> Predicted short-term (one- or two-hour average) concentrations at individual stations show larger deviations from observed concentrations than do the long-term predictions. However, over a period of a month or a season, the overall distribution of predicted short-term concentrations closely approximates the distribution of observed concentrations.

> The calm, or light wind, case is not adequately treated by the Gaussian plume type of urban diffusion model. Further study of procedures for applying the model to this type of situation is needed.

b. Study of SCIM, CDM, AQDM, and Gifford-Hanna Models

In a later study, Koch and Fisher (1973) carried out an evaluation of three Gaussian models, namely, the Sampled Chronological Input Model (SCIM), the Climatological Dispersion Model (CDM), the Air Quality Display Model AQDM), and a fourth, box, model, the simplified Gifford-Hanna Model (GHM). The SCIM generates a long-term average concentration by calculating one-hour-average concentrations for a limited number of selected hours from the required long period (e.g., one year). These one-hour-average concentrations are then averaged to obtain the long-term mean. CDM and AQDM calculate a long-term mean concentration by utilizing a distribution of meteorological conditions for the period interest. Concentrations are calculated for various meteorological conditions and weighted by their relative frequency of occurrence. The GHM model merely assumes that pollutant concentrations are directly proportional to source strengths and inversely proportional to wind speed. To ascertain the consequences of options in preparing model inputs, Koch and Fisher studied several variations of the models. In the preparation of data inputs, the emissions rates, stability conditions, and wind speed were assumed either to be constant throughout the data period or to vary from hour to hour. In addition, for the calculations with GHM, the concentrations originating from point sources were either added or not added in the calculations. In all, ten model variations were studied as follows:

(1) SCIM--variable area source emissions rates, atmospheric stability, the height of the mixing layer.

(2) SCIM--constant area source emissions rates, variable atmospheric stability and height of the mixing layer.

(3) SCIM--constant area source emissions rates, atmospheric stability, and height of the mixing layer.

(4) GHM--constant area source emissions rates and wind speed, without point sources.

(5) GHM--variable area source emissions rates and wind speed, without point sources.

(6) GHM--constant area source emissions rates and wind speed, with point sources.

(7) GHM--variable area source emissions rates and wind speed, with point sources.

(8) CDM--constant atmospheric stability and height of the mixing layer.

(9) CDM--variable atmospheric stability and height of the mixing layer.

(10) AQDM.

Two types of air quality data were used: one-hour-average SO_2 concentrations at 10 locations and annual mean SO_2 and particulate concentrations at 127 locations in the vicinity of New York City. Meteorological data were obtained from La Guardia and Kennedy Airports.

The one-hour data were used to compare SCIM and GHM. Results are given in Table A-9. They are similar to the results obtained by Koch and Thayer (1971) for short-term predictions: Correlation coefficients are < 0.45, slopes of regression of observed on predicted concentrations are different from unity, and the root-mean-square error for the individual differences is of the same order as the mean values.

As expected, the models performed best when used for predicting annual mean concentrations. These results are shown in Tables A-10 and A-11. Comparison with the data in Table A-9 shows that root-mean-square errors are lower and correlation coefficients are higher for the annual averages than for the one-hour averages. No single model stands out as being better than all of the others in every respect. Koch and Fisher (1973) reached the following conclusions based on this study:

- > The use of variable emissions rates for SCIM and GHM does not result in any conclusive improvement in model performance over the use of mean emissions rates. It is inferred that this result is due to the failure to properly treat other causes of variance, such as those associated with atmospheric stability.
- > Based on the results for New York City, the Climatological Dispersion Model (CDM) and SCIM versions of the multiple-source Gaussian plume model produce a smaller station-to-station root-mean-square error than does the Air Quality Display Model (AQDM) version (i.e., RMSEs of 52 and 59, respectively, compared with 92, with an overall mean of 135 $\mu g/m^3$ for SO_2; RMSEs of 22 and 22 compared with 36, with an overall mean of 82 $\mu g/m^3$ for particulates).
- > Although the New York City evaluation statistics for GHM, CDM, and SCIM are similar for SO_2, GHM results for particulates have a much higher station-to-station root-mean-square error than do CDM and SCIM (i.e., RMSE of 60 compared with 22, with an overall mean of 82 $\mu g/m^3$).

TABLE A-9. COMPARISONS OF MEASURED AND PREDICTED ONE-HOUR-AVERAGE SO_2 CONCENTRATIONS IN NEW YORK CITY

		NYC Station Number									
		6	1	3	10	14	17	27	28	31	36
Number of Comparisons		2892	1213	1489	1670	1641	1768	1702	1381	1783	1675
Mean (µg/m³)	Measured	304	212	296	420	230	193	188	127	187	145
	SCIM (Variable Q, S, H)	421	271	374	307	172	192	210	98	214	186
	SCIM (Mean Q, Var. S, H)	440	251	390	320	184	196	211	102	221	176
	SCIM (Mean Q, S, H)	282	157	248	224	94	102	118	90	121	100
	CHM (without points)	308	196	270	369	105	130	128	65	97	97
	CHM (with points)	330	226	313	424	131	169	156	90	186	96
Standard Deviation (µg/m³)	Measured	220	157	192	327	127	119	129	88	108	108
	SCIM (Variable Q, S, H)	421	405	388	281	223	235	255	180	297	221
	SCIM (Mean Q, Var. S, H)	421	412	406	272	243	229	252	184	281	230
	SCIM (Mean Q, S, H)	170	148	163	137	73	87	97	85	96	119
	CHM (without points)	200	125	173	236	68	84	81	41	60	38
	CHM (with points)	209	140	181	247	76	103	91	54	121	100
Root Mean Square Error (µg/m³)	SCIM (Variable Q, S, H)	400	392	400	410	216	237	251	188	295	219
	SCIM (Mean Q, Var. S, H)	455	416	433	430	238	262	268	195	285	244
	SCIM (Mean Q, S, H)	261	201	240	402	185	165	163	128	166	151
	CHM (without points)	219	165	238	350	176	142	144	109	144	147
	CHM (with points)	227	168	236	354	156	137	137	104	145	146
Mean Absolute Error (µg/m³)	SCIM (Variable Q, S, H)	233	192	228	256	146	152	146	112	139	137
	SCIM (Mean Q, Var. S, H)	280	219	254	277	160	156	159	118	166	166
	SCIM (Mean Q, S, H)	188	142	168	257	147	125	118	96	110	130
	CHM (without points)	154	111	167	207	137	103	108	79	108	111
	CHM (with points)	160	114	166	213	118	98	97	75	99	106
Maximum Measured (µg/m³)		2242	1350	1723	6711	1345	1009	1178	836	1072	903
Error at Maximum Measured (µg/m³)	SCIM (Variable Q, S, H)	994	-351	-1463	-6502	-606	-745	-867	-827	-943	-641
	SCIM (Mean Q, Var. S, H)	-258	-508	-1469	-6512	-778	-852	-934	-827	-942	-696
	SCIM (Mean Q, S, H)	-1546	-960	-1579	-6511	-1218	-953	-1052	-882	-901	-708
	CHM (without points)	-1373	-1154	-1490	-6429	-1044	-630	-966	-787	-1009	-650
	CHM (with points)	-1125	-892	-1473	-6358	-1022	-629	-902	-787	-911	-736
Correlation Coefficient	SCIM (Variable Q, S, H)	0.43	0.30	0.23	0.17	0.40	0.24	0.29	0.17	0.21	0.26
	SCIM (Mean Q, Var. S, H)	0.20	0.20	0.16	0.08	0.1[illegible]	0.15	0.14	0.12	0.18	0.19
	SCIM (Mean Q, S, H)	0.13	0.30	0.13	0.02	0.29	0.14	0.15	0.05	0.19	0.18
	CHM (without points)	0.46	0.34	0.16	0.28	0.29	0.24	0.29	0.19	0.21	0.33
	CHM (with points)	0.46	0.36	0.21	0.26	0.36	0.27	0.31	0.14	0.24	0.13
Reduction of Variance by Linear Regression	SCIM (Variable Q, S, H)	0.18	0.09	0.05	0.03	0.16	0.06	0.08	0.03	0.04	0.06
	SCIM (Mean Q, Var. S, H)	0.04	0.04	0.02	0.00	0.11	0.02	0.02	0.02	0.03	0.04
	SCIM (Mean Q, S, H)	0.02	0.04	0.02	0.00	0.08	0.02	0.02	0.00	0.04	0.03
	CHM (without points)	0.21	0.11	0.03	0.08	0.08	0.06	0.09	0.04	0.04	0.11
	CHM (with points)	0.21	0.13	0.04	0.07	0.13	0.07	0.09	0.02	0.06	0.02
Slope of Regression Line	SCIM (Variable Q, S, H)	0.22	0.12	0.11	0.20	0.23	0.12	0.15	0.08	0.06	0.12
	SCIM (Mean Q, Var. S, H)	0.10	0.08	0.08	0.04	0.17	0.08	0.07	0.08	0.07	0.08
	SCIM (Mean Q, S, H)	0.17	0.21	0.15	0.06	0.50	0.18	0.20	0.08	0.22	0.16
	CHM (without points)	0.50	0.42	0.18	0.38	0.53	0.34	0.47	0.41	0.37	1.31
	CHM (with points)	0.49	0.41	0.22	0.35	0.61	0.31	0.44	0.22	0.21	0.13
Intercept of Regression Line	SCIM (Variable Q, S, H)	210	181	254	380	191	170	157	119	171	125
	SCIM (Mean Q, Var. S, H)	258	191	267	407	198	176	172	121	172	131
	SCIM (Mean Q, S, H)	235	179	259	407	182	174	163	123	161	129
	CHM (without points)	148	150	247	279	173	148	128	100	151	89
	CHM (with points)	133	120	227	271	150	141	119	107	154	133

Source: Koch and Fisher (1973).

TABLE A-10. MODEL COMPARISONS FOR ANNUAL MEAN SO_2 CONCENTRATIONS USING NEW YORK CITY DATA

Statistic	SCIM	SCIM $(\bar{Q})$	SCIM $(\bar{Q},\bar{S},\bar{H}_m)$	GHM $(\bar{Q},\bar{U})$	GHM	GHM $(\bar{Q},\bar{U})$ + Points	GHM + Points	AQDM	CDM	CDM $(\bar{S},\bar{H}_m)$
Number of Comparisons	75	75	75	71	71	71	71	75	75	75
Mean Measured ($\mu g/m^3$)	135	135	135	140	140	140	140	135	135	135
Mean Calculated ($\mu g/m^3$)	163	162	88	78	94	107	123	211	138	206
Mean Error ($\mu g/m^3$)	28	27	-47	-62	-46	-33	-17	76	3	71
Root-Mean-Square Error ($\mu g/m^3$)	59	59	65	78	70	58	59	121	52	124
Mean Absolute Error ($\mu g/m^3$)	46	47	50	67	59	46	44	92	37	89
Largest Negative Error ($\mu g/m^3$)	-112	-106	-149	-171	-170	-139	-133	-87	-118	-112
Largest Positive Error ($\mu g/m^3$)	169	162	50	104	162	151	209	310	166	332
Error Range ($\mu g/m^3$)	281	268	200	274	332	290	342	397	284	444
Correlation Coefficient	0.84	0.83	0.82	0.82	0.82	0.83	0.83	0.89	0.84	0.84
Reduction of Variance (%)	71	69	68	67	67	70	69	79	70	71
Slope of Regression Line	0.70	0.70	0.98	0.85	0.70	0.76	0.64	0.45	0.66	0.41
Intercept of Regression Line	20	21	48	74	74	59	61	31	35	40
Maximum Measured ($\mu g/m^3$)	385	385	385	385	385	385	385	350	350	350
Error for Maximum Measured ($\mu g/m^3$)	-47	-58	-149	-75	-11	-10	53	112	-101	13

Source: Koch and Fisher (1973).

TABLE A-11. MODEL COMPARISONS FOR ANNUAL MEAN PARTICULATE CONCENTRATIONS USING NEW YORK DATA

Statistic	SCIM	SCIM $(\bar{Q})$	SCIM $(\bar{Q},\bar{S},\bar{H}_m)$	GHM $(\bar{Q},\bar{U})$	GHM	GHM $(\bar{Q},\bar{U})$ + Points	GHM + Points	AQDM	CDM	CDM $(\bar{S},\bar{H}_m)$
Number of Comparisons	114	114	114	112	112	112	112	113	113	113
Mean Measured ($\mu g/m^3$)	81	81	81	82	82	82	82	82	82	82
Mean Calculated ($\mu g/m^3$)	69	69	58	92	104	101	113	102	74	88
Mean Error ($\mu g/m^3$)	-12	-13	-24	11	22	19	31	20	-8	6
Root-Mean-Square Error ($\mu g/m^3$)	22	22	55	60	77	64	82	36	22	28
Mean Absolute Error ($\mu g/m^3$)	16	16	33	36	44	38	47	28	16	21
Largest Negative Error ($\mu g/m^3$)	-68	-66	-83	-66	-63	-61	-57	-51	-63	-60
Largest Positive Error ($\mu g/m^3$)	-43	39	463	325	405	338	419	115	68	98
Error Range ($\mu g/m^3$)	110	106	546	391	468	400	475	166	131	158
Correlation Coefficient	0.68	0.68	0.30	0.66	0.66	0.67	0.67	0.62	0.61	0.64
Reduction of Variance (%)	46	46	9	43	43	45	45	39	37	41
Slope of Regression Line	0.78	0.80	0.13	0.21	0.18	0.21	0.17	0.38	0.63	0.42
Intercept of Regression Line	28	26	74	62	63	61	62	43	35	45
Maximum Measured ($\mu g/m^3$)	169	169	169	169	169	169	169	169	169	169
Error for Maximum Measured ($\mu g/m^3$)	-54	-52	-83	150	208	161	219	5	-48	-6

Source: Koch and Fisher (1973).

Proper design and interpretation of a model evaluation study can aid in improving model performance. The results of these two studies show that the validity of an air quality model cannot necessarily be judged by a single statistic: A comprehensive program is necessary if valid conclusions are to be reached. Model evaluation statistics can vary from one station to another and from one averaging time to another, and one figure of merit cannot necessarily provide a basis for choosing one model over another. In addition, comparison of results obtained using different types of inputs was used to infer possible problems with model formulation. Moreover, in the latter study of several models, which was carried out so that direct model comparisons could be made, conclusions were drawn as to the relative performance of the various models. It is clear that if model evaluations were carried out according to standardized procedures, studies of different models by different workers would be directly comparable.

These studies had a shortcoming that is common for evaluative studies to date: The data used for the evaluation were not collected specifically for that purpose. The large amount of pollutant concentration data available, however, did allow some conclusions to be drawn about model performance.

c. Studies of CRSTER Model

Two studies were carried out in 1975 to examine the performance of the EPA single source model CRSTER (Mills and Record, 1975; Mills and Stern, 1975). The first of these studies was carried out for the Canal Power Plant, located near Cape Cod. The second studied the Stuart, Muskingum, and Philo power plants in Ohio.

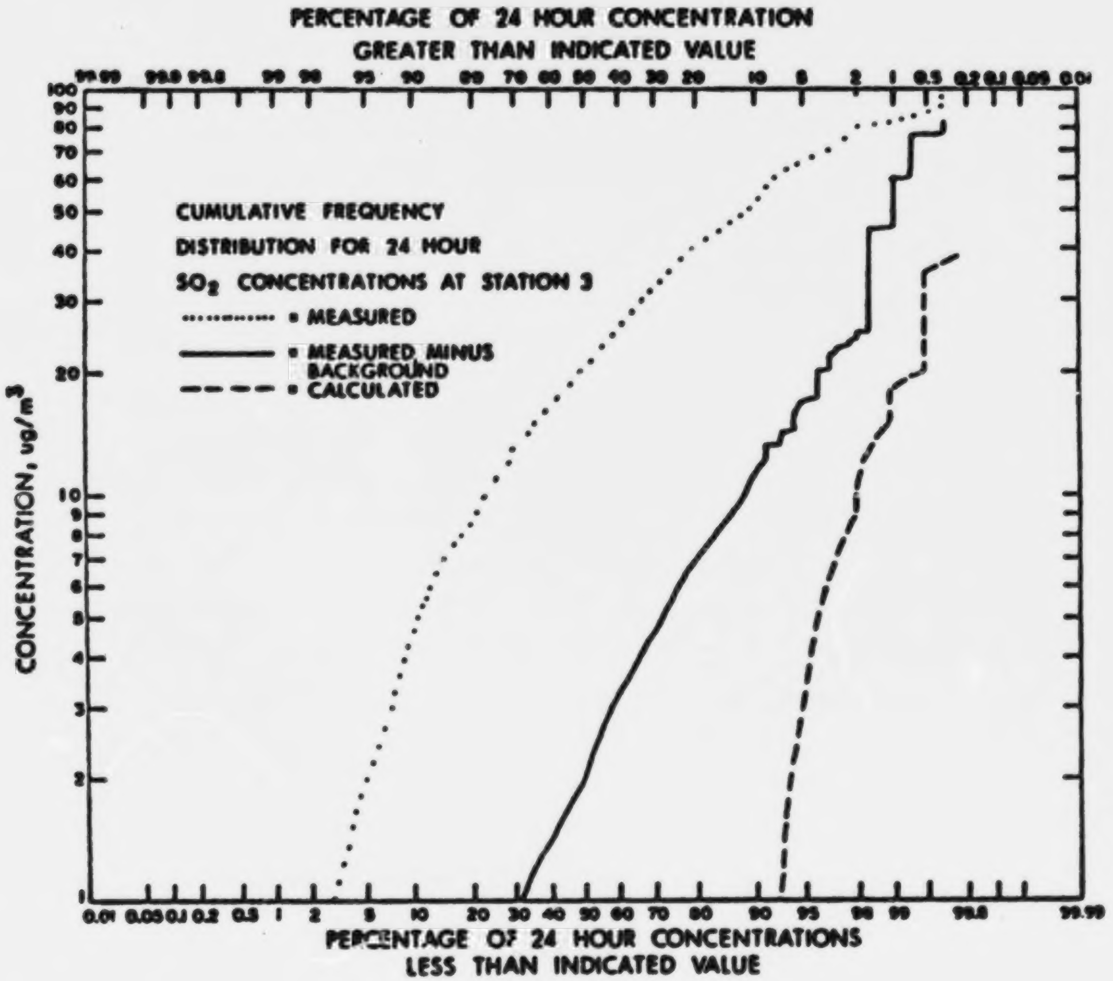
Comparisons between predicted and observed results were made by constructing frequency distributions. In the Canal study, the model consistently underpredicted measured concentrations (less so at higher concentrations), which was attributed to an overestimation of plume height by the model and a consequent underprediction of pollutant levels at the monitoring stations. In addition, Mills and Record pointed out that the calculated plume spread was not large enough to affect more than one model receptor location, which resulted in few predicted low and medium concentrations. A typical set of frequency distributions is shown in Figure A-1. No attempt was made to quantify statistically the agreement between calculated and measured distributions (e.g., by a Kolmogorov-Smirnov or χ^2 test).

In the second study (Mills and Stern, 1975), much better agreement between measured and calculated concentrations was found. Figure A-2 shows one of the frequency distributions for the Stuart plant. The better agreement between measured and calculated results at higher concentrations was attributed to the reduced influence of uncertainties associated with the determination of background concentrations. The method of determining background concentrations, which was to average the concentrations at all stations upwind from the plant using the wind direction measured at the plant, sometimes resulted in downwind concentrations less than background, producing negative net measured concentrations. Better agreement at low levels was obtained when background was added to predicted concentrations (Figure A-2), so that all concentrations (observed and predicted) were nonnegative.

In both of these studies comparison was made through the use of frequency distributions. Such comparison methods fail to associate a particular observed concentration with the corresponding computed value. Scatter plots are useful visual devices for these comparisons.

d. Study of PTMTP Model

Another Gaussian mdoel for which an evaluation study was carried out is the EPA's PTMTP Model (Guzewich and Pringle, 1977). PTMTP is used to calculate one-hour-average concentrations at a number of receptors resulting from the emissions of up to 25 point sources. The effluent characteristics of the sources (i.e., parameters employed to estimate plume rise) as well as hourly meteorological data are input to the model. In this study an inert tracer (SF_6) was injected into a spray dryer stack during



Source: Mills and Record (1975).

FIGURE A-1. A TYPICAL CUMULATIVE FREQUENCY DISTRIBUTION OF 24-HOUR-AVERAGE SO_2 CONCENTRATIONS MEASURED AND CALCULATED USING THE CRSTER MODEL

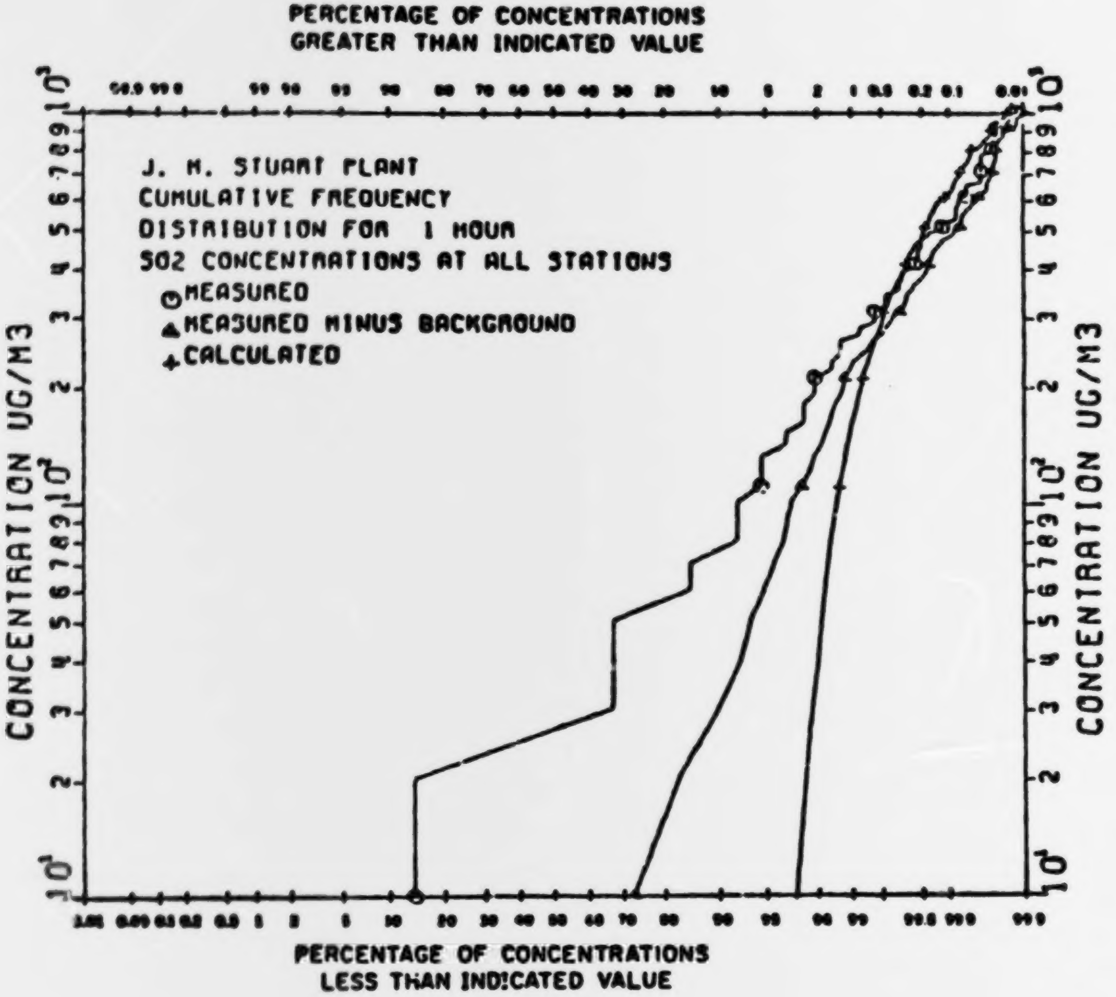
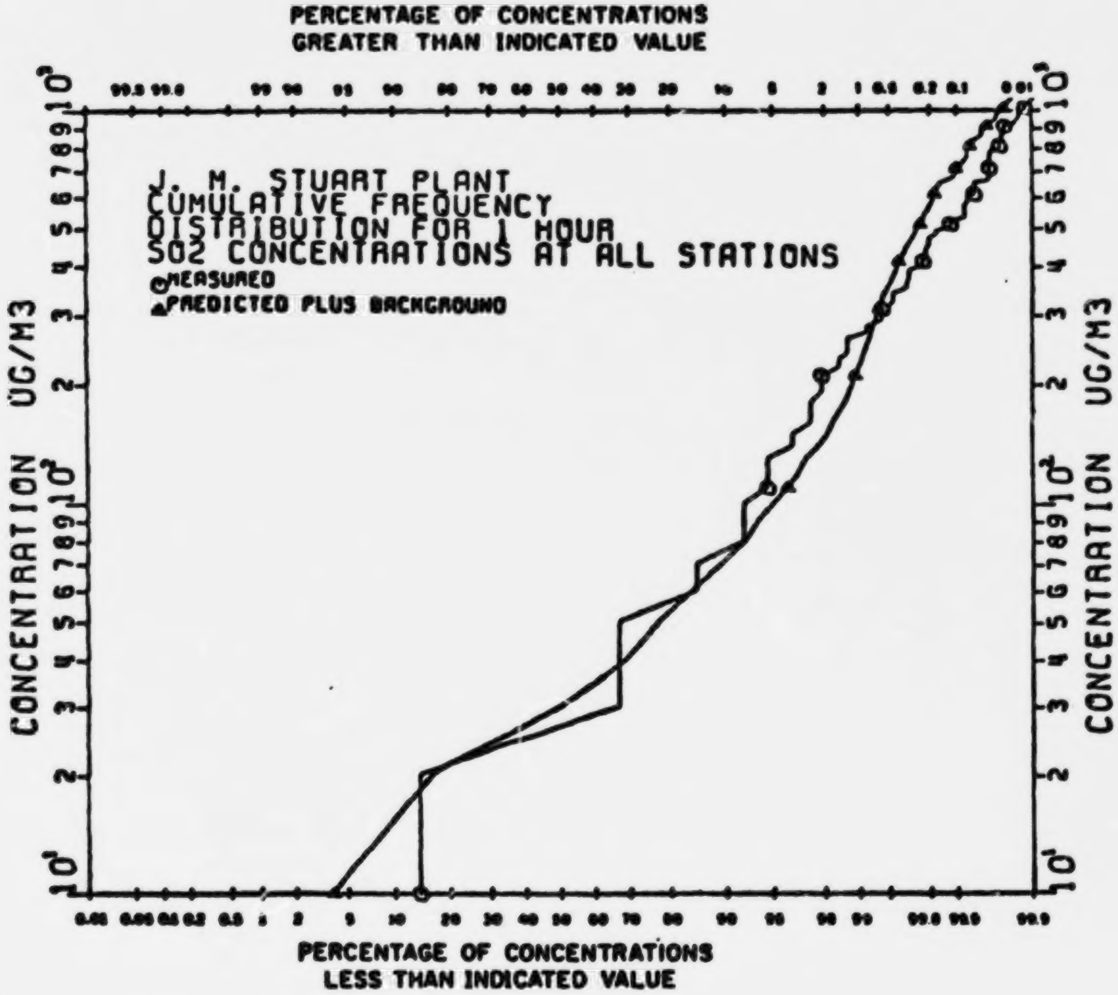


FIGURE A-2. J. M. STUART PLANT CUMULATIVE FREQUENCY DISTRIBUTION FOR ONE-HOUR-AVERAGE SO_2 CONCENTRATIONS AT ALL STATIONS. Number of measured concentrations = 45,512; number of calculated concentrations = 61,320.



Source: Mills and Stern (1975).

FIGURE A-2 (Concluded)

various meteorological conditions, and downwind tracer concentrations were monitored. The results were assessed by calculating correlation coefficients and by plotting the data. The findings are shown in Table A-12 and Figure A-3.

The data presented in Table A-12 show that good correlation was obtained for stability categories C and D, whereas very poor correlation was observed for stability categories E and F. The poor correlations were attributed to the paucity of the data (60 percent of the observations were below the detection limit of the SF_6 monitoring instrumentation) and to the resultant small data sets. Another potential source of error is the identification of the stability category. The value used for the vertical stability coefficient, σ_z, depends on the stability category, and Bohac et al. (1974) pointed out that concentrations predicted by the Gaussian plume model can be very sensitive to σ_z.

Figure A-3 shows the results of regressing the predicted on the observed data. The authors pointed out that "every predicted concentration was within a factor of 10 of the value measured and 89 percent of all predicted values were within a factor of three of the measured concentration." It may be seen from Figure A-3 that while the regression line is fairly close to the ideal, the 95 percent confidence interval on the dependent variable is of the same order of magnitude as the total range of its values.

The correlation coefficient is a statistic commonly used in evaluation work. However, although it is a useful indicator of a relationship between two data sets, it has some drawbacks, most notably that it is scale invariant. (This implies that if all numbers in one data set were doubled, the correlation coefficient would not change.) Thus it should be used in conjunction with another performance measure that can indicate the correspondence in scale between observed and predicted results. The use of regression analysis in this context has been discussed by Brier (1973, 1975), as mentioned later in this appendix.

The Gaussian plume formulation was also tested by Shum et al. (1975), who found that, for C and D stabilities, 72 percent of measured concentrations were within a factor of two of model predictions, and for B

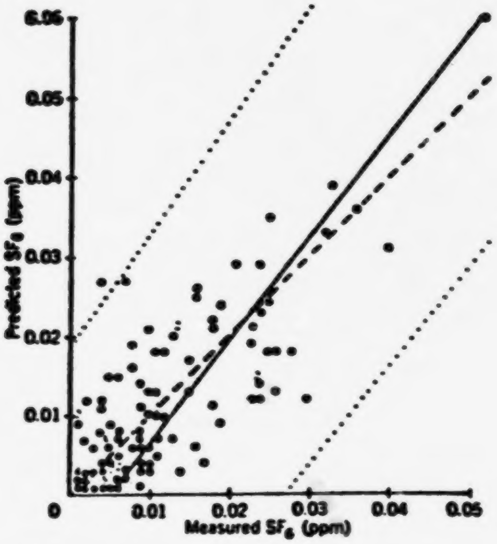
TABLE A-12. CORRELATION BETWEEN OBSERVED SF_6 CONCENTRATIONS AND CONCENTRATIONS CALCULATED USING PTMTP, BY STABILITY CLASS

	Stability Class						
Correlation	A	B	C	D	E	F	ALL
Excluding zero concentrations	-- (0)*	0.29 (19)	0.79 (33)	0.54 (45)	-0.58 (6)	-- (2)	0.77 (107)
Including zero concentrations	-- (0)	0.36 (20)	0.79 (33)	0.63 (55)	-0.1 (20)	0.46 (5)	0.81 (133)

---

* Figures in parentheses are numbers of observations.

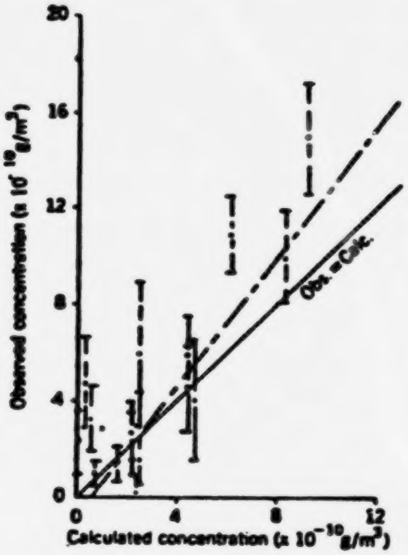
Source: Guzewich and Pringle (1977).



Source: Guzewich and Pringle (1977).

FIGURE A-3. MEASURED SF_6 CONCENTRATIONS AND PREDICTIONS USING PTMTP FOR STABILITY CATEGORIES B THROUGH F. The solid line shows the least square regression line that best fits the data; the dotted lines indicate the 95 percent confidence interval; the dashed line shows predicted concentration equals measured concentration.

stability, 63 percent were within a factor of two. Figure A-4 shows their results for C stability. It can be seen that in this case also there is a wide uncertainty interval around each point.



Source: Shum et al. (1975).

FIGURE A-4. SCATTER DIAGRAM COMPARING OBSERVED AND CALCULATED CONCENTRATIONS OBTAINED FROM THE STANDARD GAUSSIAN PLUMR MODEL UNDER SLIGHTLY UNSTABLE CONDITIONS (C STABILITY) FOR THE HIGH SOURCE AT THE WESTERN KRAFT CORPORATION

The evidently large uncertainty in the observed concentration measurements argues for some indication of the uncertainty in the parameters of the regression line (intercept, slope) to be given. Moreover, there are theoretical problems associated with the use of regression analysis in this context, as pointed out by Brier (1973). In particular, the assumption independence of data values is clearly violated by air quality observations.

e. Study of CDM Model

A more recent study of the Gaussian model CDM in the Copenhagen area was reported by Prahm and Christensen (1977). Predicted three-month-average SO_2 concentrations were compared with measurements at 24 stations covering a "flat" urban area of 500 km^2. Different combinations of parameters were used in the model runs, as follows:

(1) σ_L, dispersion parameter for area (low) sources:
  (a) original Pasquill parameters.
  (b) original Pasquill parameters, shifted one stability class toward the unstable region.
  (c) parameters due to McElroy (1969).

(2) σ_H, dispersion parameter for point (high) sources:
  (a) Hogstrom parameters (see Brummage, 1968)
  (b) Singer and Smith parameters (see Brummage, 1968)
  (c) McElroy parameters.

(3) σ_0, initial dispersion at a source:
  (a) 20 meters
  (b) 30 meters
  (c) 40 meters.

(4) $T_{1/2}$, half-life of SO_2:
  (a) 1 hour
  (b) 3 hours
  (c) x hours.

The results are shown in Table A-13. For the 24 stations, the squared correlation coefficient (r^2) varied between 0.6 and 0.7. However, data from two stations appeared as outliers in all cases. When data from these stations were excluded (they were in "questionable surroundings"), the squared correlation coefficients varied between 0.82 and 0.65. The authors stated, "thus the model explains more than 80 percent of the spatial variation of the SO_2 concentrations in the urban area." The correlation coefficient was not sensitive to the parameter combinations used, a result attributed by the authors to the large number of sources distributed at various distances from the receptor points. When measured concentrations were regressed on calculated concentrations, slopes closest to 1.0 were obtained for parameter values corresponding to rapid dilution and reaction.

TABLE A-13. PARAMETER COMBINATIONS EMPLOYED IN THE PRAHM AND CHRISTENSEN STUDY AND A SUMMARY OF THE RESULTS

	Test Number	1a	1b	2a	2b	2c	3a	3b	3c	4a	4b	5a	5b	6a	6b	7a	7b	8a	8b	9a	9b
σ_L	Pasquill a															X	X	X	X		
	Pasquill b																			X	X
	McElroy	X	X	X	X	X	X	X	X												
σ_H	Hogstrom									X	X	X	X								
	Singer Smith	X	X	X	X	X	X	X	X							X	X	X	X	X	X
	McElroy													X	X						
σ_0	20 m	X	X																		
	30 m			X	X	X				X	X			X	X	X	X			X	X
	40 m						X	X	X			X	X					X	X		
T_1	1 hour		X			X			X		X		X		X		X		X		
	3 hours	X			X			X		X		X		X		X		X		X	
	x hours			X			X														
r^1	24 stations	0.64	0.63	0.64	0.63	0.62	0.64	0.63	0.61	0.66	0.65	0.66	0.65	0.64	0.62	0.66	0.65	0.66	0.65	0.64	0.62
r^2	22 stations	0.84	0.84	0.83	0.83	0.82	0.83	0.82	0.82	0.85	0.84	0.84	0.84	0.83	0.83	0.84	0.84	0.83	0.83	0.82	0.83
r	22 stations	0.92	0.92	0.91	0.91	0.91	0.91	0.91	0.91	0.92	0.92	0.92	0.92	0.91	0.91	0.92	0.92	0.91	0.91	0.91	0.91
α	slope	0.74	0.79	0.82	0.84	0.88	0.92	0.93	0.98	0.87	0.92	0.96	1.2	0.87	0.93	0.54	0.57	0.60	0.64	0.72	0.76
β	cut-off	40	43	35	38	40	32	36	39	33	37	31	35	39	42	39	42	37	40	37	40

Source: Prahm and Christensen (1977).

This study is notable for its use of model evaluation to study the structure of the model and look at many possible parameter combinations for input. However, because of the problems with use of regression analysis, the significance of the differences in slopes and intercepts for the various cases is not clear.

f. Study of TRAPS, CALINE-2, HIWAY, and AIRPOL-4 Models

A study of four line-source models incorporating the Gaussian dispersion formula was carried out by Maldonado and Bullin (1977), using CO data from five experimental programs. This study was a part of the development activity for a new model, the Texas Roadway Air Pollution Simulator (TRAPS). Results are given in Table A-14 and Figure A-5.

When TRAPS was compared with the other three models over the complete data set, it was found to be superior over all. It had lower average error (an indication of bias), lower mean-squared error (an indication of general agreement with measured concentrations), and higher percentages of results within ±1 ppm and ±2 ppm of measured concentrations. The regression lines in Figure A-5 indicate that the TRAPS Model most closely approaches the ideal relationship, although the authors give no details of the regression statistics. These regressions show graphically that the HIWAY Model exhibited the worst fit of the predicted to the measured concentrations. HIWAY also has the highest mean-squared error for all of the data sets except one. This effect apparently was due to a small number of extremely large errors, both positive and negative, since neither the average error nor the percentages of results within ±1 ppm and ±2 ppm were consistently different for HIWAY relative to the other three models. Again, we note the use of regression analysis in model evaluation.

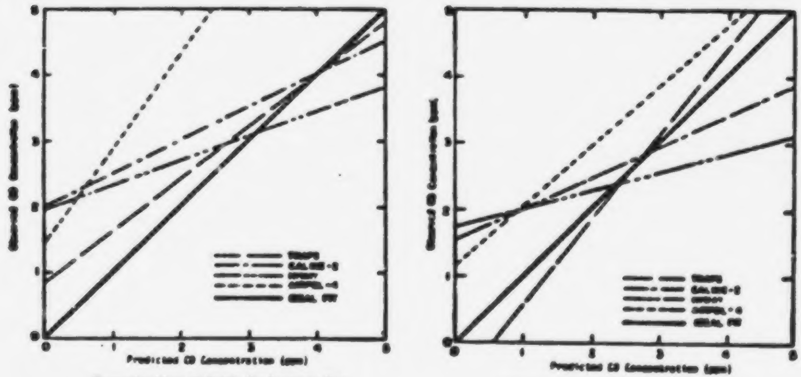
g. Study of APRAC-1A Model

A validation study of the APRAC-1A Model was carried out by Dabberdt et al. (1973). APRAC-1A is an urban diffusion model intended for predicting concentrations of inert, vehicle-generated pollutants. The model was evaluated using observations from St. Louis. Meteorological data

TABLE A-14. STATISTICAL RESULTS FROM THE MALDONADO AND BULLIN STUDY

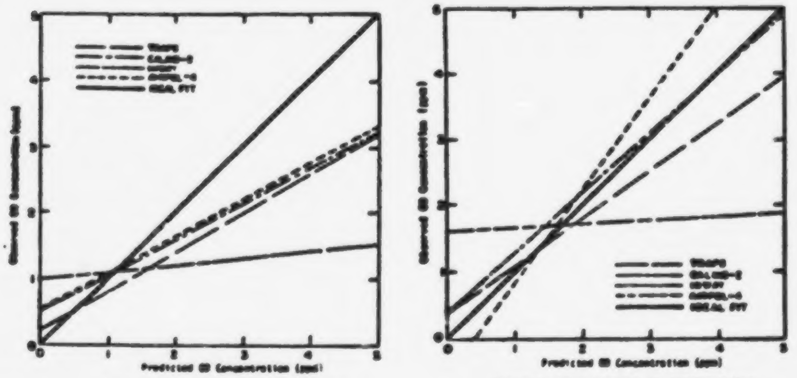
Data set	Statistic	TRAPS	CALINE-2	HIWAY	AIRPOL-4
Tennessee	No. of data pts	499	459	503	503
	Av error, ppm	0.4	1.6	1.1	1.9
	Av sqd error, ppm^2	1.1	4.0	4.0	4.7
	% within ±1 ppm	74	33	43	27
	% within ±2 ppm	95	62	76	58
North Carolina	No. of data pts	274	274	274	274
	Av error, ppm	0.0	0.2	−1.1	1.0
	Av sqd error, ppm^2	2.9	4.7	11.5	4.2
	% within ±1 ppm	48	49	45	61
	% within ±2 ppm	85	74	65	74
Virginia	No. of data pts	170	186	186	186
	Av error, ppm	−0.5	−0.1	−0.9	0.0
	Av sqd error, ppm^2	1.1	1.0	8.5	0.9
	% within ±1 ppm	74	75	67	84
	% within ±2 ppm	95	95	85	95
Illinois	No. of data pts	132	132	132	132
	Av error, ppm	0.6	1.2	−1.9	0.9
	Av sqd error, ppm^2	1.3	2.4	10.5	1.3
	% within ±1 ppm	67	41	46	60
	% within ±2 ppm	92	83	70	92
California	No. of data pts	211	211	211	74
	Av error, ppm	−0.2	0.2	−1.0	0.3
	Av sqd error, ppm^2	3.1	2.7	39.9	3.5
	% within ±1 ppm	55	62	47	56
	% within ±2 ppm	84	84	73	80
Cumulative comparison	No. of data pts	1168	1262	1306	1161
	Av error, ppm	0.1	0.8	−0.3	1.2
	Av sqd error, ppm^2	1.8	3.3	12.7	3.5
	% within ±1 ppm	65	48	48	47
	% within ±2 ppm	91	76	74	73

Source: Maldonado and Bullin (1977).



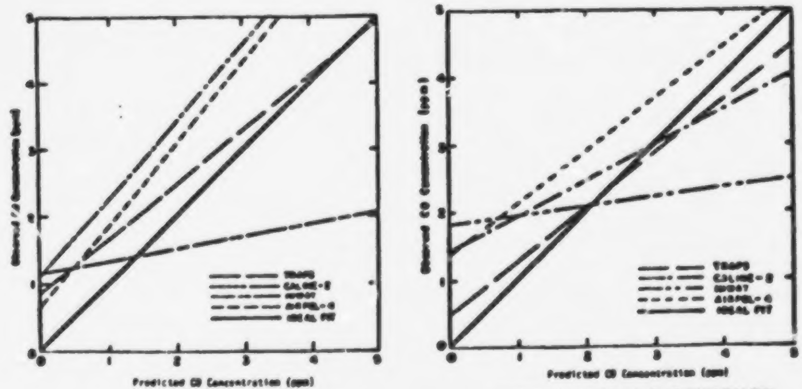
Regression Lines of Models for Tennessee Data

Regression Lines of Models for North Carolina Data



Regression Lines of Models for Virginia Data

Regression Lines of Models for California Data



Regression Lines of Models for Illinois Data

Regression Lines of Models for Combined Data Sets

Source: Maldonado and Bullin (1977).

FIGURE A-5. COMPARISON OF CO MEASUREMENTS AND PREDICTIONS FROM FOUR MODELS EXAMINED BY MALDONADO AND BULLIN were obtained from the St. Louis Airport and the NWS station at Salem, Illinois. Traffic data were obtained from the Missouri Highway Department, except for average vehicle speeds and diurnal traffic cycles measured in the downtown area. It was felt that using data that were not specially collected represented the way the model would be applied by a user and therefore would be a realistic test of the applicability and accuracy of the model.

The results are shown in Figures A-6 and A-7, which give the diurnal variations in measured and predicted CO concentrations for two weeks and the measured and predicted cumulative frequency distributions of CO concentrations. The root-mean-square difference between measured and predicted concentrations ranged from 2.6 to 3.9 ppm. Calibration of the model reduced this difference to 1.6 to 3.3 ppm. Correlation coefficients for the different locations were in the range from 0.4 to 0.7.

Of the possible sources of error in the model, Dabberdt et al. cited two as being most likely to account for the discrepancies between measurements and predictions. First, the minimum transport speed was assumed to be 1 m/s, which, it was suggested, is too low for uban areas. This error would result in a tendency to overestimate, particularly the higher concentrations. Second, the emissions submodel, which relates emissions rate to average vehicle speed over a composite test route, was probably inadequate to specify the microscale distribution of emissions. The model would probably predict average concentrations over an area better than concentrations at a particular point.

2. NUMERICAL MODELS

Having discussed several evaluation studies of Gaussian models, we now turn to studies of models that incorporate numerical solutions of the atmospheric diffusion equation. As noted earlier, these models are based on the equations of conservation of mass for each pollutant species. The focus of this section, therefore, is on models based on the solution of the following equations [see, for example, Reynolds et al. (1973)]:

$$\overbrace{\frac{\partial c_i}{\partial t} + u\,\frac{\partial c_i}{\partial x} + v\,\frac{\partial c_i}{\partial y} + w\,\frac{\partial c_i}{\partial z}}^{\text{Advection Terms}} = \overbrace{\frac{\partial}{\partial x}\left(K_H\,\frac{\partial c_i}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_H\,\frac{\partial c_i}{\partial y}\right) + \frac{\partial}{\partial z}\left(K_V\,\frac{\partial c_i}{\partial z}\right)}^{\text{Turbulent Diffusion Terms}} \tag{A-2}$$

$$+ R_i(c_1, c_2, \ldots, c_n, T) + S_i(x,t,z,t)$$

$$- W_i(x,y,z,t) \quad ,$$

$$i = 1, 2, \ldots, n \quad ,$$

where

c_i = time-averaged concentration of species i,

x, y, z = Cartesian coordinates, with z the vertical coordinate,

u, v, w = components of the wind vector in the x, y, and z directions, respectively,

t = time,

n = number of species,

K_H, K_V = horizontal and vertical turbulent diffusivities, respectively,

T = temperature,

R_i = volume rate of production of species i through chemical reactions,

S_i = rate of emission of species i from volume sources,

W_i = rate of removal of species i through scavenging mechanisms.

These equations form the basis of all the numerical models considered in this section. Solution of the equations in the form given can be extremely complex; gathering, preparing, and supplying the requisite input data can also be tedious and time-consuming. Thus, simplifying assumptions are frequently made, leading to more easily solvable forms of the equation or to reduced requirements for input information. To some extent, there is a trade-off between the degree of simplification on the one hand and the expected accuracy and reliability of prediction on the other. Since these models are formulated in a fundamentally different manner from Gaussian models, they may present different problems in evaluation.

The numerical models described herein fall into three categories: grid, trajectory, and box models. In formulations of grid models, the region of interest is divided into a three-dimensional array of "cells," each perhaps 1 to 4 kilometers on a side and on the order of 10 to several hundred meters high. In the trajectory approach, a hypothetical column of air advected by the wind is followed through the modeling region. The box model, which is conceptually the simplest, treats the entire region of interest as a well-mixed cell.

Since grid models predict pollutant concentrations averaged over a complete cell, the problem of representativeness mentioned earlier becomes an issue. Care should be exercised in the placing of pollutant monitors for collecting comparison data so that the point measurement taken by the monitor is representative of average concentration over an area the size of a grid cell. This problem is of importance for both primary and secondary pollutants. For example, CO and NO concentrations in the immediate vicinity of a source can deviate significantly from the spatially averaged values. For secondary pollutants such as NO_2 and O_3, some time must elapse between the release of their precursors and their ultimate formation through photochemical reactions, and this time allows for the pollutant cloud to become more spatially homogeneous. Nevertheless, microscale phenomena affect the concentrations of these pollutants as, for example, in the depletion of O_3 in the vicinity of a roadway by fresh emissions of NO.

Liu et al. (1976a) reported three previous evaluation studies:

> "Further Development and Evaluation of a Simulation Model for Estimating Ground Level Concentrations of Photochemical Pollutants," R73-19, Systems Applications, Incorporated, Beverly Hills (now in San Rafael), California (February 1973).
> "Evaluation of a Diffusion Model of Photochemical Smog Simulation," EPA-R4-73-012, Volume A (CR-1-273), General Research Corporation, Santa Barbara, California (October 1972).
> "Controlled Evaluation of the Reactive Environmental Simulation Model (REM)," EPA-R4-73-013a, Volume I, Pacific Environmental Services, Incorporated, Santa Monica, California (February 1973).

The models included were the SAI Airshed Model, the GRC model DIFKIN, and the PES model REM. The SAI model is a grid model; the other two are trajectory models. Data for four pollutants from six measurement stations in the Los Angeles basin for six smoggy days in late summer and early fall of 1969 were used for comparison. The performance measures used to evaluate the models were:

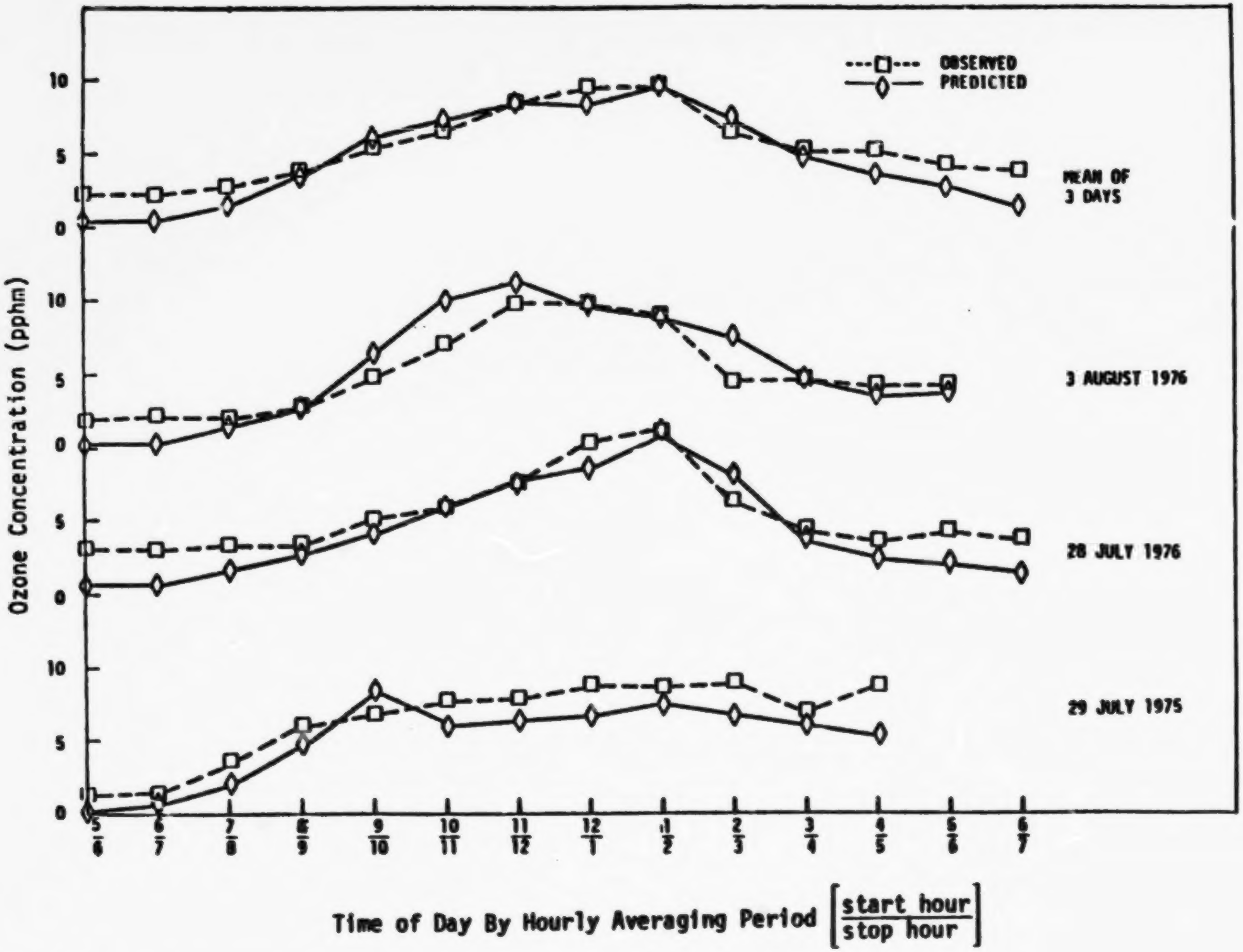
> Correlation coefficients.
> Root-mean-square deviation between measurements and predictions.
> χ^2 test on the residuals, comparing them with a normal distribution.
> Scatter plots of predictions versus measurements.
> Plots of residuals against
  - Time of day
  - Predicted concentrations
  - Measured concentrations.
> Histograms of residuals.

The above set of evaluation methods was chosen to detect both random and systematic failure of the models to account for the observed results. However, Liu et al. commented, "the results of statistical tests are relatively insensitive indicators of model performance because of the limited quantity of data, the varying conditions and assumptions, the nondistributional character of the data, and the complexity of the potential source of error. One should not substitute statistical analysis results for an examination of the plots."

The correlation coefficients were mostly between 0.5 and 0.9, which is generally higher than results achieved by Gaussian models for one-hour averages. Correlations were higher for CO and NO, both primary pollutants, than they were for NO_2 and O_3, which are secondary pollutants.

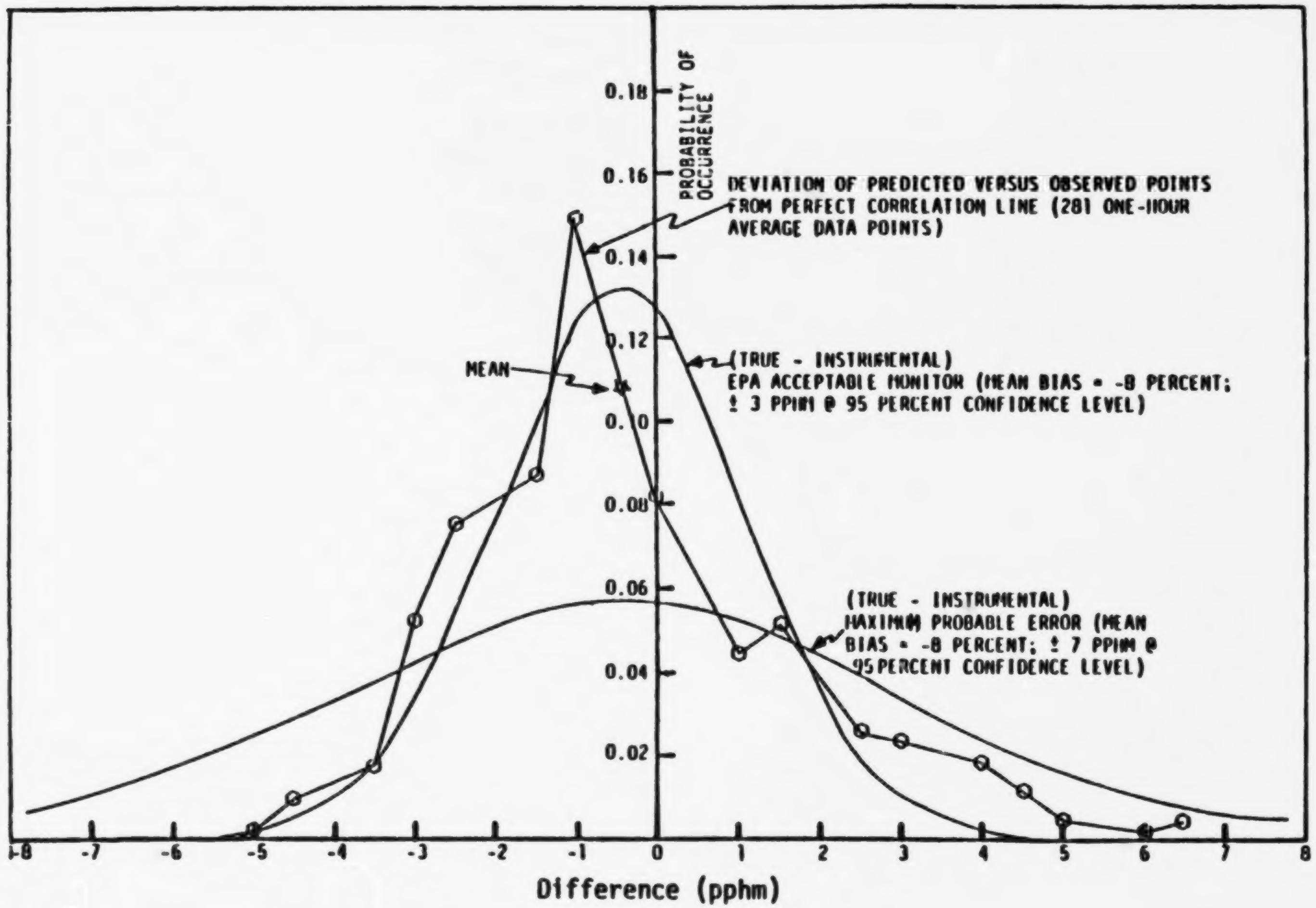
The main conclusion from the study was that none of the three models could be said to have been adequately validated, mainly because of the sparseness and unrepresentativeness of the data base. This illustrates the point that evaluation studies utilizing existing data are less likely to reach satisfactory conclusions as are studies for which the data-gathering effort is an integral part of the evaluation effort. In addition, we note that these efforts represent an early stage in the development of photochemical models; a considerable amount of developmental work has been carried out subsequent to these studies.

A validation study of a later version of the SAI Airshed Model was carried out by Anderson et al. (1977). Predictions of ozone concentrations were compared with measurements for three days in 1975-1976 at nine Denver, Colorado, monitoring stations. The average time variations in predicted and measured ozone concentrations are shown in Figure A-6. Note that the predictions follow measurements fairly closely. It was also shown in this study that the residuals compared very closely in both mean and standard deviation with the expected error distribution of a measuring instrument (Figure A-7). The authors concluded that the observational evidence is not precise enough to establish confidence limits on the model predictions. Overall, the correspondence of predictions to measurements is very close, both for concentrations above the NAAQS for ozone and for all of the data (see Table A-15).



Source: Anderson et al. (1977).

FIGURE A-6. TIME VARIATIONS OVER ALL STATIONS OF OBSERVED ONE-HOUR-AVERAGE OZONE CONCENTRATIONS AND THE CORRESPONDING PREDICTIONS OBTAINED FROM THE SAI AIRSHED MODEL



Source: Anderson et al. (1977).

FIGURE A-7. PREDICTIONS OF THE SAI AIRSHED MODEL COMPARED WITH ESTIMATES OF INSTRUMENT ERRORS FOR OZONE (DATA FOR 3 DAYS, 9 STATIONS, DAYLIGHT HOURS)

TABLE A-15. OCCURRENCE OF CORRESPONDENCE LEVELS OF PREDICTED AND OBSERVED OZONE

(percentage of comparisons meeting correspondence level)

Correspondence Level Between Predicted and Observed Pairs	Comparisons	Both Predicted and Observed Concentrations > 8 pphm
Factor of 2 (2P > O > P/2)	80%	94%
Computed value is within ± twice standard deviation maximum probable instrument error (95% level) of observed value	100	100
Computed value is within ± standard deviation of maximum probable instrument error (95% level) of observed value	93	90
Computed value is within ± twice standard deviation of instrument errors by EPA standard (95% level) of observed value	89	77
Computed value is within ± standard deviation of instrument errors by EPA standard (95% level) of observed value	60	37

Source: Anderson et al. (1977).

CO, NO, and NO_2 concentrations were predicted less well. Discrepancies in the CO results were attributed to microscale effects not included in the model's formulation. Observational data for NO and NO_2 were insufficient to evaluate the model.

Anderson et al. (1977) drew the following conclusions from their study of the SAI Airshed Model:

- The model is a very good predictor of one-hour-average ozone concentrations in grid cells in the Denver region.
- The model's ozone predictions at any given station probably have at least as narrow an error distribution as do measurements at the same station.
- If predictions and measurements are equally accurate, model predictions can be expected to be within a factor of 2 of true concentrations 80 percent of the time, and its predictions of exceedances of the NAAQS for ozone at that time (i.e., more than 8 pphm) could be expected to be within that factor 94 percent of the time.
- The accuracy of model predictions of regional maximum ozone concentrations should exceed the accuracy of model predictions of concentrations at specific stations.

A verification of the Livermore Region Air Quality Model (LIRAQ), (MacCracken et al., 1977) was carried out by MacCracken and his coworkers (MacCracken and Sauter, 1975). LIRAQ is a two-dimensional grid model (i.e., it has no vertical resolution). This was a very comprehensive study, covering data for CO, HC, NO_2, NO, and O_3 for two days in 1973. Many different evaluation statistics were evaluated, as listed in Table A-16. Results from the two study days are shown in Table A-17. Generally, the temporal correlation coefficients are higher than the station correlations, which indicates that the model follows the temporal trends in pollutant concentrations better than the spatial trends. As pointed out above, this effect is to be expected in a grid model, since predicted concentrations are averaged over relatively large grid squares and the

TABLE A-16. STATISTICAL MEASURES USED IN VERIFICATION OF THE LIRAQ PHOTOCHEMICAL MODEL

Statistical Measure	Definition
Mean R_{time}	Mean correlation coefficient in time, given by: $\sum_i m_i R_i / \sum_i m_i$. where R_i is the correlation coefficient for measurement vs predictions based on one-hour-averaged station records at the ith station, and m_i is the number of one-hour-average records at the ith station.
Median R_{time}	Median of the correlation coefficients developed for each of the stations.
$R_{station}{}^{means}$	Correlation coefficient for the measured and predicted mean concentrations at the stations, where the averaging is over only those hours during the air quality simulation for which measurements exists.
$R_{station}{}^{maxima}$	Correlation coefficient for the measured and predicted maximum concentrations at the various stations.
$\langle M \rangle / \langle O \rangle$	Ratio of the average predicted to the average measured concentration, where the average is over all one-hour station periods within the simulation period for which measurements exist.
$\langle M_{max} \rangle / \langle O_{max} \rangle$	Ratio of the average of the predicted maximum hourly concentration at each of the stations to the average of the measured maximum one-hour-average concentrations at the stations.
$\langle RMS \rangle$	Root-mean-square deviation between predicted and measured one-hour-average station records, based on all of the observed data.
R	Correlation coefficient between predictions and measurements based on all of the one-hour-average concentration measurements. (Of all the correlation coefficients calculated, only R is based on a sample size large enough to be used without a substantial correction for the degrees of freedom.)

Source: MacCracken and Sauter (1975).

MICROFILMED FROM
BEST AVAILABLE COPY

TABLE A-17. STATISTICS FOR THE LIRAQ MODEL EVALUATION STUDY

Pollutant	Statistic	Value 26 July 1973	 20 August 1973
CO	Mean R_{time}	0.67	0.35
	Median R_{time}	0.70	0.47
	$R_{station}$ means	0.58	0.59
	$R_{station}$ maxima	0.69	0.86
	$\langle M\rangle/\langle O\rangle$	0.76	1.0
	$\langle M_{max}\rangle/\langle O_{max}\rangle$	0.86	0.81
	$\langle RMS\rangle$	1.5	0.77
	R	0.56	0.58
HC	Mean R_{time}	0.62	0.54
	Median R_{time}	0.74	0.56
	$R_{station}$ means	0.43	0.62
	$R_{station}$ maxima	0.25	0.74
	$\langle M\rangle/\langle O\rangle$	0.80	0.79
	$\langle M_{max}\rangle/\langle O_{max}\rangle$	0.89	0.78
	$\langle RMS\rangle$	1.4	1.00
	R	0.54	0.60
NO_2	Mean R_{time}	0.44	0.33
	Median R_{time}	0.79	0.50
	$R_{station}$ mean	0.48	0.33
	$R_{station}$ maxima	0.28	0.30
	$\langle M\rangle/\langle O\rangle$	0.88	1.31
	$\langle M_{max}\rangle/\langle O_{max}\rangle$	1.04	1.12
	$\langle RMS\rangle$	3.13	1.76
	R	0.54	0.34
NO	Mean R_{time}	0.71	0.64
	Median R_{time}	0.80	0.53
	$R_{station}$ means	0.85	0.37
	$R_{station}$ maxima	0.46	0.86
	$\langle M\rangle/\langle O\rangle$	1.82	0.91
	$\langle M_{max}\rangle/\langle O_{max}\rangle$	1.65	0.87
	$\langle RMS\rangle$	6.1	1.95
	R	0.72	0.62
O_3	Mean R_{time}	0.79	0.74
	Median R_{time}	0.84	0.86
	$R_{station}$ means	0.68	0.62
	$R_{station}$ maxima	0.79	0.64
	$\langle M\rangle/\langle O\rangle$	0.85	1.56
	$\langle M_{max}\rangle/\langle O_{max}\rangle$	0.80	1.4
	$\langle RMS\rangle$	2.4	1.94
	R	0.78	0.73

Source: MacCracken and Sauter (1975).

measurements with which they are compared are taken at particular points within grid squares. The RMS errors vary widely, but in the absence of any indication of the accuracy or precision of the instruments used to obtain the measurements, it is not possible to assess the significance of these figures.

The authors concluded that for 26 July 1973 the model gave a fair overall fit to most measurements and a very good fit to the O_3 data. The model also gave a good representation of temporal trends for that day. It was also concluded that, though NO was overpredicted, other species were slightly underpredicted, a result that the authors said was expected considering the 5 km resolution of the model.

The August 1973 day studied by MacCracken and Sauter (1975) was characterized by strong winds and low pollutant concentrations. It was concluded that the spatial and temporal correlations for this day were quite good for all species except NO_2, which was overpredicted quite markedly. This overprediction was attributed to too high a boundary concentration above the mixing layer.

MacCracken and Sauter (1975) pointed out that statistical verification measures can be heavily influenced by outlying points. For instance, the statistic <M>/<O>* can be substantially diminished by one large measurement (particularly if few results are being compared). They pointed out that examination of concentration patterns as well as statistics can alleviate this problem. In addition, the use of statistics that are relatively insensitive to outlying data values can help to minimize their effects (see for example, the study by Cleveland et al., 1976). Care must be taken, however, that the data values whose importance is being diminished are truly spurious, and not actual high concentrations that the model is failing to predict correctly.

The SAI Reactive Plume Model (RPM) was evaluated in the course of a study to determine the feasibility of ozone formation in power plant plumes (Tesche et al., 1976). The results from 16 field experiments were used to assess the performance of RPM. Data obtained from aircraft flights operated by Meteorology Research, Incorporated, and the University of Washington provided the data base for initial conditions, and, farther downwind, concentration data for comparison with model predictions.

---

* The ratio of the average predicted to the average measured concentration (see Table A-16).

The dilution scheme used by RPM was examined using data from a previous study with an inert tracer (SF_6) (Liu et al., 1976b). Good agreement between RPM predictions and tracer measurements was observed, as evaluated visually from data plots, suggesting that the transport and dispersion portion of RPM performs quite satisfactorily.

Comparisons of measured and predicted concentrations of reactive species were made for NO_x, O_3, and SO_2. In the case of NO_x, there were discrepancies between the measurements by the two aircrafts, and the model predictions agreed more closely with the MRI results. For both NO_x and O_3, the model tended to overpredict. The discrepancies were attributed to uncertainty in the hydrocarbon measurements needed for the chemical reaction calculations. Later study showed that some hydrocarbon concentrations used as input to RPM were too high, which would have had the effect of inflating the predicted concentrations of both NO_2 and O_3. Moreover, a sensitivity study showed that the variable to which RPM predictions were most sensitive was the ambient reactive hydrocarbon concentration. Another source of uncertainty in the model predictions was the background pollutant concentrations, which were obtained from the aircraft traverses of the plume. This information was used to prescribe the concentrations of ambient pollutants entrained into the plume. The aircraft measurements were taken at particular locations and times, however, and cannot completely describe the possible temporal and spatial variations in the background reactive hydrocarbon concentration field.

The overall conclusions of the study, based on graphical comparisons of measured and predicted concentrations, were:

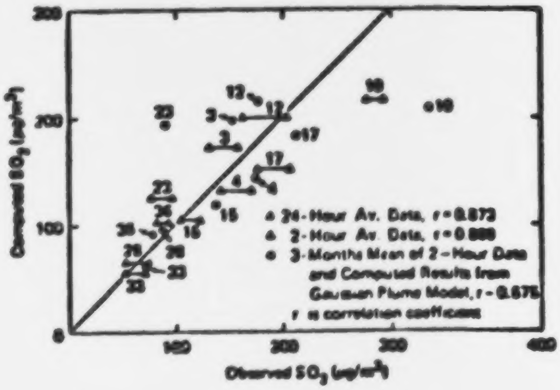
> The dilution of conservative pollutants and tracer material is described well by RPM.

> NO_x and O_3 were generally overpredicted in the study, probably in large part because of excessively high estimates of entrained hydrocarbons.

> The accuracy of the model for ozone predictions at large downwind distances is of the order of ±25 percent. This accuracy is commensurate with the accuracy of the field data used.

A numerical grid model developed by Shir and Shieh (1974) was evaluated as part of its development, with data for a 25-day period in St. Louis during February 1965. The model was developed to calculate SO_2 concentrations, and only first-order chemical reaction effects were considered. The air quality data base employed in the study consisted of observations from 10 stations that reported two-hour-averaged concentrations. In addition, there were instruments at those stations that recorded the 24-hour-average concentrations.

Figure A-8 shows a comparison of measured and predicted concentrations averaged over the 25 days. Correlation coefficients of 0.873 for the 24-hour-average data and 0.899 for the two-hour-average data were obtained. No explanation was offered by Shir and Shieh (1974) as to why the correlation coefficient for two-hour averages was higher than that for 24-hour averages. Presumably this discrepancy is an artifact of the different instruments used for the measurements.



Source: Shir and Shieh (1974).

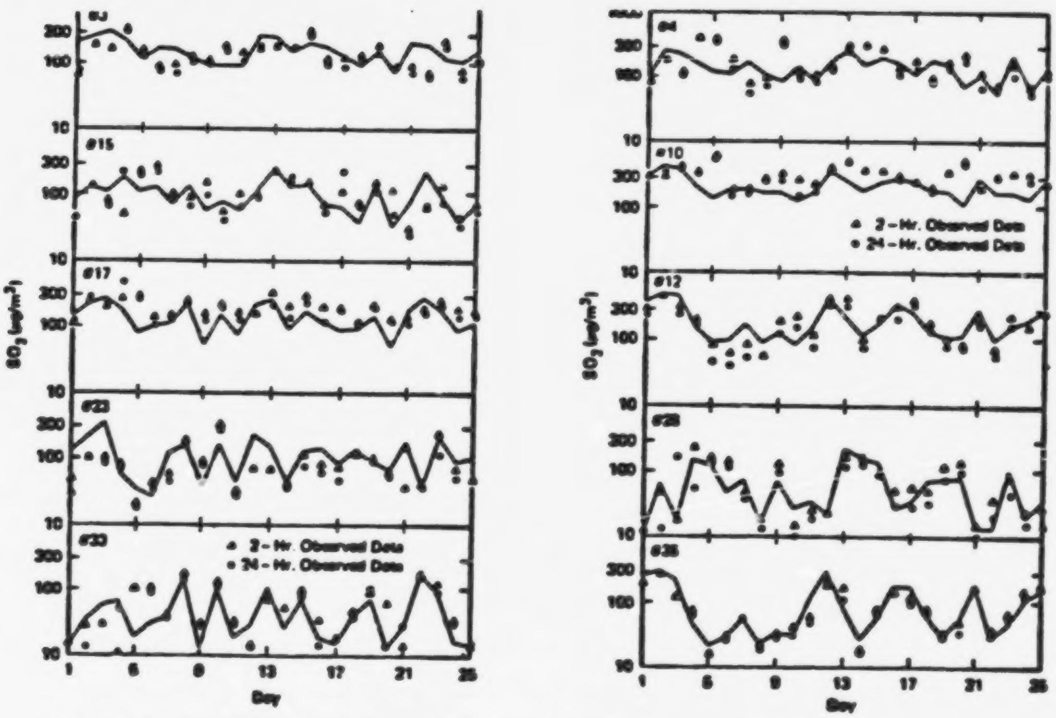
FIGURE A-8. COMPARISON OF MEASURED AND PREDICTED 25-DAY-AVERAGED SO_2 CONCENTRATIONS (1-26 FEBRUARY 1965)

Figure A-9 shows the comparison between measured and predicted 24-hour-average concentrations; Figure A-9 (a) shows the temporal variations of the daily values, and Figure A-9 (b) illustrates the results in a set of scatter plots. Discrepancies were ascribed to (1) underestimated emissions rates from the emissions inventory model, and (2) microscale effects that cannot be modeled on the grid scale used. The scatter plots show good agreement for most stations, and the overall correlation coefficients between measured and predicted values are 0.806 for the logarithms of the concentrations and 0.654 for the concentrations themselves.

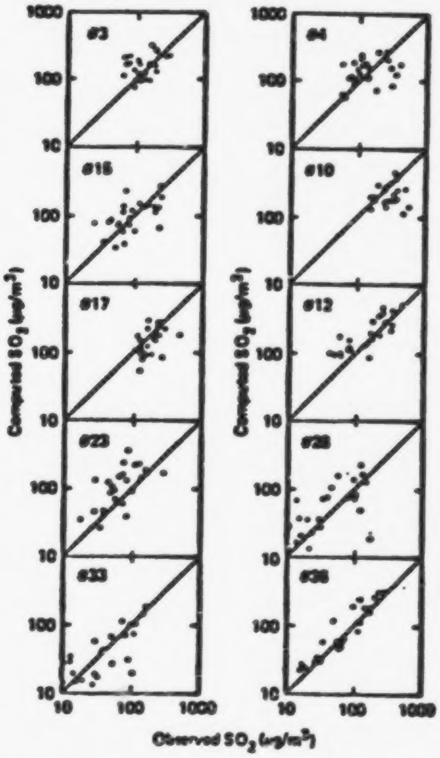
Figure A-10 shows the comparison between observed and predicted two-hour-average concentrations. The Shir and Shieh model predicts these shorter-term averages less well, as may be seen by comparing Figures A-9 and A-10. The correlations for the results in Figure A-10 were 0.706 for the logarithms of the concentrations and 0.531 for the concentrations themselves. We note that these correlation coefficients are lower than those calculated for the 24-hour-average concentrations, in contrast to the station-averaged data presented above. Again, no explanation was advanced by the authors for this discrepancy.

The frequency distributions for the two-hour-average concentrations were then classified according to different variables and compared. As illustrated in Figure A-11, pollutant frequency distributions were constructed based upon the prevailing wind direction.

Comparisons for four different wind sectors are shown. The agreement is best for the NE and SE wind directions. For NW winds, concentrations were overpredicted, and for SW winds, they were underpredicted. The authors hypothesized that since these winds often carry cold air and warm air, respectively, the discrepancies may be due to lag response of the emissions rates of changing air temperatures. This hypothesis was checked by constructing the frequency distributions for different temperature ranges, shown in Figure A-12. Although the agreement is good for temperatures in the range $-3 < T < 3°C$, concentrations are overpredicted for temperatures below -3°C and underpredicted for temperatures above 3°C.



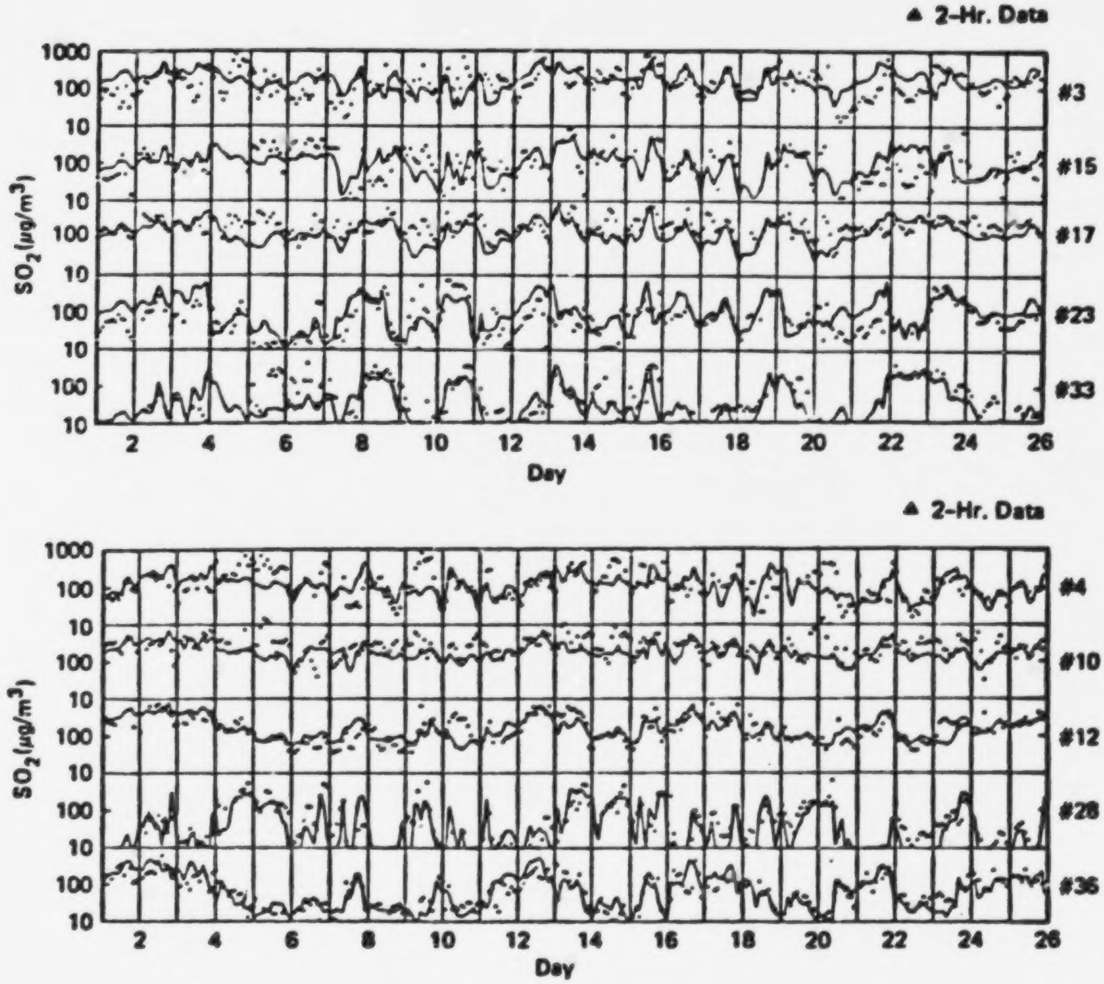
(a) 24-Hour-Average Variations of SO_2 Concentrations for a 25-Day Period at Each Monitoring Station



(b) 24-Hour-Average SO_2 Concentrations at Each Station

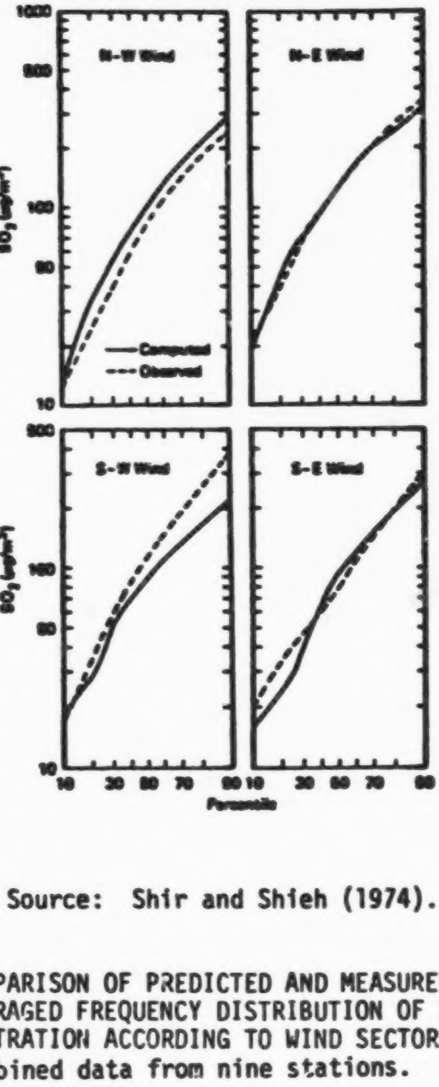
Source: Shir and Shieh (1974).

FIGURE A-9 COMPARISON OF PREDICTIONS AND MEASUREMENTS OF SO_2 CONCENTRATIONS IN ST. LOUIS REPORTED BY SHIR AND SHIEH



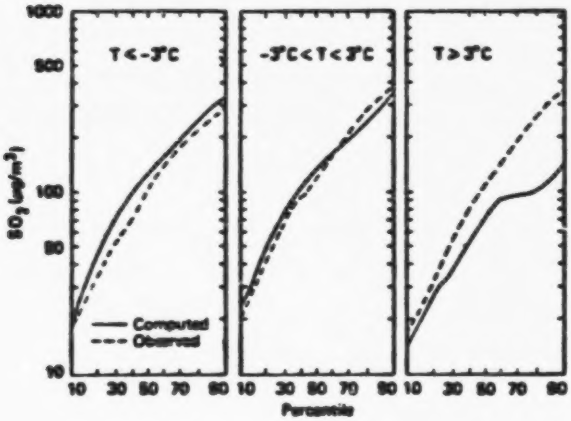
Source: Shir and Shieh (1974).

FIGURE A-10. COMPARISON OF PREDICTED AND MEASURED TWO-HOUR AVERAGED SO_2 CONCENTRATIONS AT EACH MONITORING STATION IN ST. LOUIS. Dots represent measurements; lines represent predictions.



Source: Shir and Shieh (1974).

FIGURE A-11. COMPARISON OF PREDICTED AND MEASURED TWO-HOUR-AVERAGED FREQUENCY DISTRIBUTION OF SO_2 CONCENTRATION ACCORDING TO WIND SECTORS. Combined data from nine stations.



Source: Shir and Shieh (1974).

FIGURE A-12. COMPARISON OF FREQUENCY DISTRIBUTIONS OF PREDICTED AND MEASURED TWO-HOUR-AVERAGE SO_2 CONCENTRATIONS ACCORDING TO AMBIENT TEMPERATURE RANGE

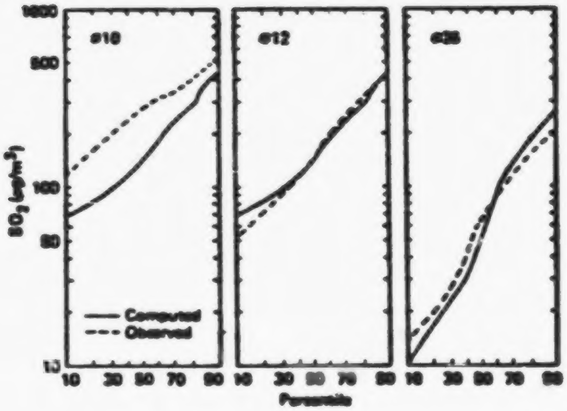
In Figure A-13, three station frequency distributions are shown. Good agreement is found for two of the stations, but at Station 10, which is near point sources, concentrations are underpredicted.

In conclusion, the authors cited the following advantages of the model:

> Consistent performance under different conditions, particularly in both strong and light winds, and for sudden shifts in wind.

> Flexibility in handling arbitrary distributions of sources with different emissions rates, spatial and temporal variations in the wind field, eddy diffusion coefficients, stability, and mixing height.

> Ability to deal with surface roughness, topographical features, and chemical reactions (although only a limited treatment of chemistry was included in the model tested).



Source: Shir and Shieh (1974).

FIGURE A-13. COMPARISON OF FREQUENCY DISTRIBUTIONS OF PREDICTED AND MEASURED TWO-HOUR-AVERAGE SO_2 CONCENTRATIONS FOR THREE STATIONS

The following disadvantages were pointed out:

> Neglect of microscale effects.

> Possible problems with the numerical integration methods; these could be solved by using more accurate methods.

> Neglect of the turbulence attributable to the effect of the urban area.

> Inaccurate representation of temperature dependence of emissions.

> Lack of representativeness of the data from monitoring stations.

The studies cited here indicate that evaluation studies of numerical models have, in general, been more thorough and comprehensive than studies of Gaussian models. This is probably a result of their comprehensive nature as models; these models are applied to multiple pollutants with chemical reactions and cover a complete urban area. Thus, a numerical model uses comprehensive inputs and produces much output, requiring a fairly comprehensive study to evaluate its various aspects.

3. MODEL EVALUATION METHODOLOGY

In the above discussion, we reviewed many model evaluation studies. Obviously much effort has been directed towards evaluating air quality models. However, relatively little work has addressed the considerations to be taken into account when designing such a study--the appropriate level and methods of data collection and suitable performance measures and standards. We now discuss three studies that deal with the methodology of model evaluation. Brier (1973) examined the often-used practice of "calibrating" models with observed data. Calibration consists of calculating a linear regression of observed on calculated values for a given set of station measurements and then using this derived relationship to convert future computed values to "true" concentrations. Brier concluded that calibration "is not a statistically valid procedure when used to make predictions of air quality for a distribution of emissions differing from that under which the calibration was actually established." Brier based this conclusion on two considerations:

- The data sample for which the calibration relationship is derived does not correspond to the situation to which it is later applied. In statistical terms, the population from which the relationship is derived is not the same as the one to which it is applied.
- Use of the regression procedure entails the assumption that the data points are statistically independent. The data used in calibration do not satisfy this assumption because of spatial and temporal correlations of concentrations.

Brier suggested that, if no alternative to calibration were available, the procedure might be made more meaningful by:

> Improvement of measurement and sampling techniques.
> Making separate calibrations for different sets of conditions, e.g., for different meteorological situations.
> Use of different statistical models.

In a later study, Brier (1975) examined some statistical questions relating to comparisons between predicted and measured values. He discussed the relationships among errors in model inputs, errors due to imperfection in the model, and errors in the observations used for validation. Suggestions were made as to estimating separately the input and output data uncertainties and the errors introduced by model inadequacies. However, the methods required some assumptions about independence of error sources, and Brier concluded that more work was necessary to see if a solution to this problem could be found. Finally, recommendations on model evaluation procedures were made. First, Brier stated that "it does not seem desirable to specify a fixed set of rules to be followed blindly under all conditions. However, certain general guidelines and suggestions can be provided that should be applicable in most cases to give assistance in planning and executing a validation study." A sensitivity analysis of the model under study was suggested as an important requirement for verification. The purposes of the sensitivity analysis in this context would be:

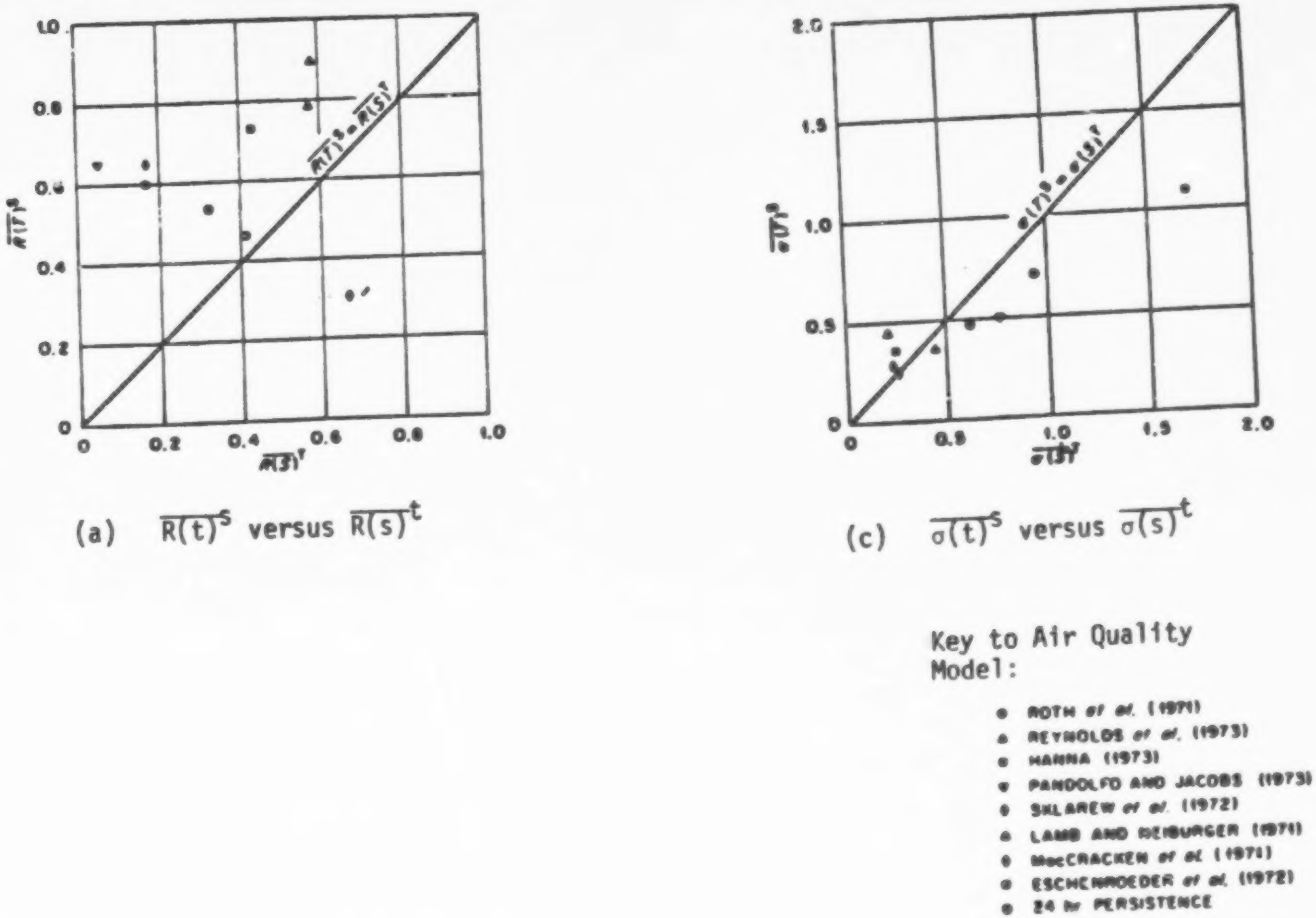
> To reveal internal inconsistencies in the model.
> To identify the parameters that dominate the model's operation.
> To provide guidance for data collection.
> To investigate error propagation through the model.

After the sensitivity analysis is complete, further analysis would center on comparisons between observed and predicted concentrations. Statistics such as correlation coefficients, variances of observed and predicted results, mean square errors, mean absolute differences, statistics of the regression of observed on predicted concentrations, and characteristics of the error distribution were suggested for evaluation. Attention should be given to departures from normal distributions in evaluating some of these statistics. Brier suggested that all of these results should be considered in evaluating the model, since they can all provide information on different aspects of the comparison between observed and predicted concentrations.

Nappo (1974) also studied methods for evaluating air quality models. He pointed out that judging the relative merits of different models that embody both temporal and spatial resolution on the basis of temporal correlation coefficients at various stations is inadequate. He also pointed out that such models should be judged not only on how well they followed temporal trends, but on the accuracies of their spatial predictions. He defined the quantities $\overline{R(t)}^s$, the average over all monitoring stations of the temporal correlation coefficients, and $\overline{R(s)}^t$, the time-averaging of the spatial correlation coefficients. Ideally, both quantities should be equal to 1, and the model's overall quality is to be judged by how closely this goal is approached. Figure A-14(a) shows a plot of these quantities for nine air quality simulation models. Eight of these models show better temporal correlations than spatial correlations. Thus, conclusions drawn on the basis of one of these criteria are not necessarily valid when the other criterion is considered. In addition to the model's ability to follow trends in the data, measured by the correlation coefficient, Nappo pointed out that an important feature of performance is the model's ability to predict the correct quantities of pollutants formed. This ability is measured by the terms $\overline{r(t)}^s$, the space average of the time-averaged ratios of predicted to measured concentrations, and $\overline{r(s)}^t$, the time average of the space-averaged ratios. Since these quantities both represent the grand average of all ratios, they are equal, but their standard deviations are not. Denoting the respective standard deviations by $\sigma\overline{(t)}^s$ and $\sigma\overline{(s)}^t$, he derived the plot shown in Figure A-14(b). Examination of the data in that figure shows that the models considered vary widely in the certainty with which they predict the amounts of pollutant at a particular station as opposed to at a particular time. Nappo contended that only by considering all aspects of a model's performance can a true picture of its utility be obtained and can different models be adequately compared.

BLANK PAGE

BLANK PAGE



(a) $\overline{R(t)}^s$ versus $\overline{R(s)}^t$

(c) $\overline{\sigma(t)}^s$ versus $\overline{\sigma(s)}^t$

Key to Air Quality Model:

Source: Nappo (1974).

FIGURE A-14. STATISTICAL QUANTITIES CALCULATED FOR NINE AIR QUALITY SIMULATION MODELS. Quantities defined in text.

Although the work reviewed here addresses some statistical aspects of measuring model performance, it is evident that there has been a lack of studies directed toward the development of air quality model evaluation methodology. The material presented in the main body of this report represents a first step to rectify this situation.

BLANK PAGE

REFERENCES

Anderson, G.E., et al. (1977), "Air Quality in the Denver Metropolitan Region, 1974-2000," EPA-908/1-77-002, Environmental Protection Agency Region VIII, Denver, Colorado.

Argonne National Laboratory (1976), "Report to U.S. EPA of the Specialists Conference on the EPA Modeling Guideline," Energy and Environmental Systems Division, Argonne, Illinois.

Bohac, R. L., et al. (1974), "Sensitivity of the Gaussian Plume Model," Atmos. Environ., Vol. 8, p. 291

Brier, G. W. (1973), "Validity of the Air Qaulity Display Model Calibration," EPA-R4-73-017, Environmental Protection Agency, Research Triangle Park, North Carolina.

Brummage, K. G. (1968), "The Calculation of Atmospheric Dispersion from a Stack," Atmos. Environ., Vol. 2, pp. 197-224.

Burton, S.C., et al. (1976), "Oxidant/Ozone Ambient Measurement Methods: An Assessment and Evaluation," EF76-111R, Systems Applications, Incorporated, San Rafael, California.

Cleveland, W. S., et al. (1976), "Robust Statistical Methods and Photochemical Air Pollution Data," J. Air Pollut. Control. Assoc., Vol. 26, pp. 36-38.

Dabberdt, W. F., et al. (1973), "Validation and Application of an Urban Diffusion Model for Vehicular Pollutants," Atmos. Environ., Vol. 7, p. 603.

Environmental Protection Agency [EPA] (1978a) "Guideline on Air Quality Models, " EPA-450/2-78-027, Research Triangle Park, North Carolina.

__________(1978b), "Guidelines for Air Quality Maintenance Planning and Analysis, Volume 9 (Revised): Evaluating Indirect Sources," EPA-450/4-78-001, Research Triangle Park, North Carolina.

__________(1976) "Quality Assurance Handbook for Air Pollution Measurement," EPA-600/9-76-005, Research Triangle Park, North Carolina.

__________(1973), "Guide for Compiling a Comprehensive Emission Inventory (Revised)," Research Triangle Park, North Carolina.

__________ (1972), "Compilation of Air Pollutant Emission Factors," AP-42, Research Triangle Park, North Carolina.

Eschenroeder, A. Q., J. R. Martinez, and R. A. Nordsieck (1972), "Evaluation of a Diffusion Model for Photochemical Smog Simulation," CR-1-273, EPA-R4-73-012a, General Research Corporation, Santa Barbara, California.

Guzewich, D. C., and W.J.B. Pringle (1977), "Validation of the EPA-PTMTP Short-Term Gaussian Dispersion Model," J. Air Pollut. Control Assoc., Vol. 27, p. 540.

Hanna, S. R. (1973), "Urban Air Pollution Models--Why?" ATDL Contribution File No. 83, Atmospheric Turbulence and Diffusion Laboratory, Oak Ridge, Tennessee.

Hayes, S. R. (1979), "Performance Measures and Standards for Air Quality Simulation Models," EF78-93R, Systems Applications, Incorporated, San Rafael, California.

Hayes, R. S., S. D. Reynolds, and P. M. Roth (1977), "A Commentary on the Analysis of Control Measures Required to Achieve Compliance with the National Ambient Air Quality Standards: The Selection of Models and the Specification of Data Requirements," Systems Applications, Incorporated, San Rafael, California.

Hilst, G. R. (1978), "Plume Model Validation," EA-917-SY, Workshop WS-78-99, Electric Power Research Institute, Palo Alto, California.

Hougland, E. S. and N. T. Stephens (1976), "Air Pollutant Monitor Siting by Analytical Techniques," J. Air Pollut. Control Assoc., Vol. 26, p. 51.

Koch, R. C., and G. E. Fisher (1973), "Evaluation of the Multiple Source Gaussian Plume Diffusion Model--Phase I," EPA-650/4-75-018a, Environmental Protection Agency, Research Triangle Park, North Carolina.

Koch, R. C., and S. D. Thayer (1971), "Validation and Sensitivity Analysis of the Gaussian Plume Multiple-Source Urban Diffusion Model," EF-60, GEOMET, Incorporated, Gaithersburg, Maryland.

Lamb, R. G., and M. Neiburger (1971), "An Interim Version of a Generalized Urban Air Pollution Model," Atmos. Environ., Vol. 5, pp. 239-264.

Lawrence Berkeley Laboratory (1976), "Instrumentation for Environmental Monitoring," LBL-1, University of California, Berkeley, California.

Lenhard, R. W. (1970), "Accuracy of Radiosonde Temperature and Pressure--Height Determination, "Bull. Am. Meteorol. Soc., Vol. 51, pp. 842-846.

Liu, M. K., Stewart, D. A., and Roth, P. M. (1978), "An Improved Version of the Reactive Plume Model (RPM-II), 9th International Technical Meeting on Air Pollution Modeling and its Application, 28-31 August, 1978, Toronto, Canada.

Liu, M. K., et al. (1977), "Development of a Methodology for Designing Carbon Monoxide Monitoring Networks," EPA-600/4-77-019, Systems Applications, Incorporated, San Rafael, California.

__________ (1976a), "Continued Research in Mesoscale Air Pollution Simulation Modeling: Volume I--Assessment of Prior Model Evaluation Studies and Analysis of Model Validity and Sensitivity," EPA-600/4-76-016a, Systems Applications, Incorporated, San Rafael, California.

__________ (1976b), "The Chemistry, Dispersion, and Transport of Air Pollutants Emitted from Fossil Fuel Power Plants in California: Data Analysis and Emission Impact Model," EF76-18, Systems Applications, Incorporated, San Rafael, California.

MacCracken, M. C., and G. D. Sauter, eds. (1975), "Development of an Air Pollution Model for the San Francisco Bay Area," Appendix 12-3, UCRL-51920, Lawrence Livermore Laboratory, University of California, Livermore, California.

MacCracken, M. C. et al. (1977), "The Livermore Regional Air Quality Model: I. Concept and Development," UCRL-77475 Pt. 1, Rev. 2, Lawrence Livermore Laboratory, University of California, Livermore, California.

__________ (1971), "Development of a Multibox Air Pollution Model and Initial Verification for the San Francisco Bay Area," UCRL-73348, Lawrence Radiation Laboratory, Livermore, California.

MacCready, P. B. and H. R. Jex (1963), "Response Characteristics and Application Techniques of Some Meteorological Sensors," Meteorology Research, Inc., Altadena, California, and Systems Technology, Inc., Inglewood, California.

Maldonado, C., and J. A. Bullin (1977), "Modeling Carbon Monoxide Dispersion from Roadways," Environ. Sci. Technol., Vol. 11, p. 1071.

Mazarella, D.A. (1972), "An Inventory of Specifications for Wind Measuring Instruments," Bull. Am. Meteorol. Soc., Vol. 53, No. 9.

McElroy, J. L. (1969), "A Comparative Study of Urban and Rural Dispersion," J. Appl. Meteorol., Vol. 8, pp. 19-31.

Meteorology Research Inc. (1975), Technical information courtesy of T. B. Smith.

Mills, M. T., and F. A. Record (1975), "Comprehensive Analysis of Time-Concentration Relationships and Validation of a Single-Source Dispersion Model, " EPA-450/3-75-083, Environmental Protection Agency, Research Triangle Park, North Carolina.

Mills, M. T., and R. W. Stern (1975), "Model Validation and Time-Concentration Analysis of Three Power Plants," EPA-450/3-76-002, Environmental Protection Agency, Research Triangle Park, North Carolina.

Nappo, C. J., Jr. (1974), "A Method for Evaluating the Accuracy of Air Pollution Prediction Models," in Preprints of the Symposium on Atmospheric Diffusion and Air Pollution, 9-13 September 1974, Santa Barbara, California (sponsored by American Meteorological Society, Boston, Massachusetts).

Pandolfo, J. P., and C. A. Jacobs (1973), "Tests of an Urban Meteorological Pollutant Model Using CO Validation Data in the Los Angeles Metropolitan Area," Vol. 1, EPA-R4-730-025a, The Center for the Environment and Man, Incorporated, Hartford, Connecticut.

Pooler, F., Jr. (1974), "Network Requirements for the St. Louis Regional Air Pollution Study" J. Air Pollut. Control Assoc., Vol. 24, p. 228.

Prahm, L. P., and M. Christensen (1977), "Validation of a Multiple Source Gaussian Air Quality Model," Atmos. Environ., Vol. 11, pp. 791-795.

Reynolds, S. D., et al. (1973), "Further Development and Evaluation of a Simulation Model for Estimating Ground Level Concentrations of Photochemical Pollutants," EPA-68-02-0339, Systems Applications, Incorporated, San Rafael, California.

Roth, P. M., et al. (1975), "An Examination of the Accuracy and Adequacy of Air Quality Models and Monitoring Data for Use in Assessing the Impact Of EPA Significant Deterioration Regulations on Energy Developments," EF75-58R, Systems Applications, Incorporated, San Rafael, California.

_________ (1971), "Development of a Simulation Model for Estimating Ground Level Concentrations of Photochemical Pollutants," 71-SAI-21, Systems Applications, Incorporated, San Rafael, California.

Rubin, E. S. (1974), "The Influence of Annual Meterological Variations on Regional Air Pollution Modeling: A Case Study of Allegheny County Pennsylvania," J. Air Pollut. Control Assoc., Vol. 24, p. 349.

Seinfeld, J. H. (1972),"Optimal Location of Pollutant Monitoring Stations in an Airshed," Atmos. Environ., Vol. 6, p. 847.

Seinfeld, J. H. (1977), "Current Air Quality Simulation Model Utility," Department of Chemical Engineering, California Institute of Technology, Pasadena, California.

Shir, C. C., and L. J. Shieh (1974), "A Generalized Urban Air Pollution Model and Its Application to the Study of SO_2 Distributions in the St. Louis Metrolopitan Area," J. Appl. Meteorol., Vol. 13, p. 185.

Shum, Y. S., et al. (1975), "The Use of Artifical Activable Trace Elements to Monitor Pollutant Source Strengths and Dispersal Patterns," J. Air Pollut. Control Assoc., Vol. 25, p. 1123.

Sklarew, R. C., et al. (1971), "A Particle-in-Cell Method for Numerical Solution of the Atmospheric Diffusion Equation, and Applications to Air Pollution Problems," Final Report 3SR-844, Systems, Science and Software, La Jolla, California.

Tesche, T. W. (1978), "Evaluating Simple Oxidant Prediction Methods Using Complex Photochemical Models," Monthly Technical Progress Narrative No. 1, EM78-14, Systems Applications, Incorporated, San Rafael, California.

Tesche, T. W., et al. (1976), "Determination of the Feasibility of Ozone Formation in Power Plant Plumes," EA-307, Electric Power Research Institute, Palo Alto, California.

Trijonis, J. C., and K. W. Arledge (1975), "Utility of Reactivity Criteria in Organic Emission Control Strategies for Los Angeles," TRW Environmental Services, Redondo Beach, California.

U.S. Army Signal Missile Support Agency (1960), "A Comparison between the Double-Theodolite and Single-Theodolite Wind Measuring Systems," Progress Report NR-11, Wind Effect on the Aerobee, White Sands Missile Range, White Sands, New Mexico.

Weather Measure Corporation (1974), Technical Information, P.O. Box 41257, Sacramento, California 95841.

TECHNICAL REPORT DATA

(Please read Instructions on the reverse before completing)

1. REPORT NO.
EPA-450/4-79-033

2.

3. RECIPIENT'S ACCESSION NO.

4. TITLE AND SUBTITLE
Procedures for Evaluating the Performance of Air Quality Simulation Models

5. REPORT DATE
October 1979

6. PERFORMING ORGANIZATION CODE

7. AUTHOR(S)
M. J. Hillyer, S. D. Reynolds, P. M. Roth

8. PERFORMING ORGANIZATION REPORT NO.

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Systems Applications, Incorporated
950 Northgate Drive
San Rafael, California 94903

10. PROGRAM ELEMENT NO.

11. CONTRACT/GRANT NO.
68-02-2593

12. SPONSORING AGENCY NAME AND ADDRESS
Office of Air Quality Planning and Standards
U. S. Environmental Protection Agency
Research Triangle Park, North Carolina 27711

13. TYPE OF REPORT AND PERIOD COVERED

14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

16. ABSTRACT

Currently there are no standardized guidelines for evaluating the performance of air quality simulation models. In this report, we develop a procedural framework for objectively evaluating model performance. In carrying out this work, we have:

- Reviewed previous model evaluation studies.
- Developed a general procedural framework for performing an evaluation study.
- Provided specific guidance, to the extent possible, with respect to the work required in each step of the performance evaluation procedure.
- Identified gaps in present knowledge that limited our ability to provide more detailed guidance in this report, and presented recommendations for further work that will help to fill those gaps.

Because model evaluation has received relatively little systematic attention to date, we were able to identify several areas ripe for future investigation. The performance of these suggested studies will be essential to the success of the guidelines presented herein.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Air Pollution Turbulent Diffusion Mathematical Models Computer Models Atmospheric Models	Dispersion Air Quality Simulation Models Model Validation Model Evaluation	

18. DISTRIBUTION STATEMENT Release unlimited	19. SECURITY CLASS (This Report) None	21. NO. OF PAGES 159
	20. SECURITY CLASS (This page) None	22. PRICE

EPA Form 2220-1 (9-73)

146